

# **Configuration Manual**

MSc Research Project Data Analytics

Aidan Browne Student ID: 16140818

School of Computing National College of Ireland

Supervisor: Dr. Catherine Mulwa

# National College of Ireland



## **MSc Project Submission Sheet**

### School of Computing

Student Name:	Aidan Browne			
Student ID:	16140818			
Programme:	MSc Data Analytics	<b>Year:</b> 2020		
Module:	Research Project			
Lecturer:	Dr Catherine Mulwa			
Submission Due Date:	28/09/2020			
Project Title:	Fine-Grained Sentiment Anal using Deep Learning Models	ysis of Yelp Reviews		

**Word Count:** 4,866

.....

Page Count: 93

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

<u>ALL</u> internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signatu	re Alden Berne
:	
Date:	28/09/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	
Attach a Moodle submission receipt of the online	
copies).	

You must ensure that you retain a HARD COPY of the	
<b>project</b> , both for your own reference and in case a project	
is lost or mislaid. It is not sufficient to keep a copy on	
computer.	

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if	
applicable):	

# **Configuration Manual**

# Aidan Browne Student ID: 16140818

# **1** Introduction

The Configuration Manual describes the steps taken to successfully produce the results achieved by the 6 models.

# 2 Hardware Specification

The research was carried out on a Lenovo T440 laptop with the following device and windows operating system specification.

Device specifications				
Device name	DESKTOP-NH7BU6V			
Processor	Intel(R) Core(TM) i5-4300U CPU @ 1.90GHz 2.49 GHz			
Installed RAM	8.00 GB (7.69 GB usable)			
Device ID	055B10A8-4220-4889-A257-FE71117CCE4B			
Product ID	00331-20020-00000-AA903			
System type	64-bit operating system, x64-based processor			
Pen and touch	Touch support with 2 touch points			

### Figure 1: Laptop Specifications

Windows specifications			
Edition	Windows 10 Pro		
Version	1903		
Installed on	29/08/2019		
OS build 18362.959			

#### Figure 2: Windows Specification

# **3** Software and Data Installation

To successfully implement the objectives of the project the following data, packages and software were installed on the candidate's laptop.

# 3.1 Data Selection

For the purpose of the research undertaken data was downloaded from the Yelp Open Data website<sup>1</sup> to candidate's local drive. The data is provided in Java Script Object Notation (JSON) files and covers a subset of Yelp's business, review and user data. The dataset is provided for academic or non-commercial purposes and was updated on the 21st February 2020. Prior to February 2020 the data was used as part of the Yelp Data Challenge with the 13<sup>th</sup> round finishing on 31<sup>st</sup> December 2019. The challenge was an annual event for students with a top prize of \$5,000. Compressed the dataset contains over 4 gigabytes (GB) of data and over 9 GB uncompressed. There were 5 JSON files contained in the file but for the research carried out only the Business, Review and User were used. Each Business, Review and User had a unique identifier which enables the files to be merged when creating the two datasets. The period the data covered was between 20/01/2007 - 31/01/2019. Numerical attributes such as review count or review star rating can be used as features for machine learning algorithms. Non-numerical attributes such as review text or list of friends were furthered processed to be able to be used for the research. The business dataset contains features such as location, attributes and category and covered 10 metropolitan areas. The review dataset contains review text, star rating and votes received if a review is useful, funny and cool. The user dataset had attributes pertaining to the user's social network and all the user's metadata. A description of the three JSON files can be found in below table

Table 1: Yelp Open Dataset Descrip	tion
------------------------------------	------

Dataset	Record Count	Attribute Count
Business	209,393	14
Review	8,021,222	9
User	1,968,703	22

The followings steps were taken to download the JSON files.

<sup>&</sup>lt;sup>1</sup> <u>https://www.yelp.com/dataset</u>

1. Enter the highlighted information to agree to the Yelp Open Dataset license agreement before downloading

🔆 Yelp Dataset 🔷	× +				
↔ ↔ ↔ ↔	🖲 🔒 https://www.yelp.com/dataset/download				⊠ ☆
	yelp&	Dataset	Dataset	Documentation	
		Cownload Yelp Dataset ease fill out your information to download the dataset. We do not store this data nor will we use this ed it to ensure you've read and have agreed to the Dataset License ur Name ease sign by entering your initials aver read and agree to the Dataset License Download	data to email you	I, WE	

Figure 3: Yelp Open Dataset License

## 3.2 Azure Databricks

Azure Databricks was chosen to create the two datasets as it offers distributing computing provided by Databricks combined with Microsoft Azure's cloud storage capabilities. This was essential as the total size of the three JSON files used as part of research project totalled 9.6 gigabytes (GB). The largest of these was the Review (6.2GB) which could not be processed by the candidate's laptop. The following steps were followed to set up an Azure Databricks account, create storage containers for the Business, Review and User JSON files and create an Azure Key Vault to access the Azure Blob container from Databricks.

1. Set up an account with your college credentials via the Azure website<sup>2</sup>

Treate your Azure free account X	+	
(← → ୯ 🏠	🕖 🔓 https://zure.microsoft.com/en-us/free/search/76ef.jd=EAlalQobChMI4piY8bjj6QlVau7tCh32OwtoEAAYASAAEgLsofD_BWEGs&OCID=AID2000084_SEM_EAlalQobChMI4piY8bjj6QlVau7tCh32OwtoEAAYASAAEgL	
	Microsoft Azure	Ê
	Create your Azure free account today	
	Get started with 12 months of free services	
	Start free >	
	Or buy now >	
	👔 Harrier Hanney Paraman 🔂 Standard	
	$+ \qquad \qquad$	
	Territoria di annanati di annanati di annanati di anna di	
	Nave type Latitization	
	Constant Al-Constant And States	
	Al Kinacharan Aciti Agginariya Tanàn ago	
	advantage print de la Parte Partiere La Partiere La Partiere La Partiere La Partiere La Partiere La Partiere Partiere La Partiere Partiere La Partiere Parti	
	Navagatar	
	🕈 Samadalaman (10) Annana magana 🖬 Atamanan 🖸 Di Samanan	
	Tauly	
	Security Langer and Apparent Security Security Langer and Security Security Langer and	
	scaring from Minocolf Inflastrators Inflastrators Undergrand for Ser	
	What do I get?	
	With your Azure free account, you get all of this—and you won't be charged until you choose to upgrade.	
	12 months \$200 modit Aluque free	🖙 Chat with Sales 🖉

Figure 4: Azure Set UP

<sup>&</sup>lt;sup>2</sup> <u>https://azure.microsoft.com/en-gb/free/</u>

2. After you have created a free Student Account to unlock the necessary products to be used in the project a Pay-As-You-Go (PAYG) subscriptions needs to be added to the account from the offers page<sup>3</sup>.



Figure 5: Pay-As-You-Go Subscriptions (1 of 2)

A Home - Microsoft Azure × +					- σ ×
← → C û D A https://portal.az	ure.com/?quickstart=true#home				··· 🖂 🏠 👘 🖶 🖶
Microsoft Azure	P: Search reso	irces, services, and docs (G+/)		21	10 © ? © ×16140818@student.nci @
	Azure services	Il resources Subscriptions HOmsphr clusters	Resource Storage groups accounts	Virtual App Services More services	
	Recent resources			Last Manual	
	Pav-As-You-Go	Sub	icription	17 minutes ago	
	🕈 Azure for Students	Sub-	cription	26 minutes ago	
	Navigate				
	Tools Microsoft Learn C*	Azure Monitor	All resources	Cost Management	
	Learn Azure with free online training from Microsoft	Monitor your apps and infrastructure	<ul> <li>Secure your apps and infrastructure</li> </ul>	Analyze and optimize your cloud spend for free	
	Useful links Technical Documentation C <sup>*</sup> Azure Migration Tools	Azure Services (2) Find an Azure expert	Recent Azure Updates 🖻 Guickstart Center	Azure mobile app	

Figure 6: Pay-As-You-Go Subscriptions (2 of 2)

3. After the PAYG subscription has been added a Databricks resource needs to be added to your subscription. From the main portal page click Azure Databricks

<sup>&</sup>lt;sup>3</sup> <u>https://azure.microsoft.com/en-us/offers/ms-azr-0003p/</u>

A Home - Microsoft Azure × + ← → C ☆ ♡ ♣ https://portal.	zure.com/?puscatart=true#home				- 0 ×
Microsoft Azure	P. Search reso	secure, services, and door (G+y)		51	타 다 ⓒ ? ⓒ #16140818@student.nci @
	Azure services	All resources Subscriptions Holinsight dusters	Resource groups scounts	Virtual App Services More services	
	Recent resources	Тур		Last Viewed	
	Pay-As-You-Go     Azure for Students -	Subscription Subscription		17 minutes ago 26 minutes ago	
	Navigate				
	Subscriptions	Resource groups	All resources	Dashboard	
	Tools				
	Learn Azure with free online training from Microsoft	Azure Monitor Monitor your apps and infrastructure	Security Center Secure your apps and infrastructure	Cost Management Analyze and optimize your cloud spend for free	
	Useful links			Azure mobile app	
	Technical Documentation 😅 Azure Migration Tools	Azure Services 3* Find an Azure expert	Recent Azure Updates of Quickstart Center	App Store Google Play	

Figure 7: Databricks Set-up

4. Click Add

Azure Databricks - Microsoft A × +											- 6	•
	https://portal.azure.com/#blade/HubsE	xtension/BrowseR	esourceBlade/resourceType/Microsoft.	Databricks%2	Fworkspaces			120%		⊠ ☆ ±	III\ 🗆 🛎 😁 📴	-
Microsoft Azure	P Search resources	services, and do	xcs (G+/)				6				x16140818@student.nc NATIONAL COLLEGE OF IREL/	
Azure Databricks	Cefresh    O Assign tags see a subscription? Open Directory +	ubscription sett	lines									×
Filter by name	All subscriptions	~	All resource groups	~	All locations	~	All tags		$\sim$	No groupi	ng	~
1 items □ Name ↑↓	Туре ↑↓		Resource group ↑↓		Location ↑↓			Subscri	ption 1.	L.		

Figure 8: Databricks Set-up

5. Update below information creating new Resource Group, Workspace name, choosing appropriate location and choosing Pricing Tier. When updated click Review + Create

🔥 Azure Databricks Service - Mic 🗡	+	- 0 ×
(←) ୯ ଇ	🛛 🚔 https://portal.azure.com/?quickstart=true#create/Microsoft.Databricks	
Microsoft Azure	P. Search resources, sarvices, and docs (G+7)	EI 💀 D 🐵 ? 🙄 x16140818@student.nd. 🕥
Home > New >		
Azure Databricks Se	ervice	×
*Basics Networking Tags	Review + Create	
Project Details		
Select the subscription to manage de manage all your resources.	eployed resources and costs. Use resource groups like folders to organize and	
Subscription * 💿	Pay-As-You-Go	
Resource group * ③	(New) MscData V Create new	
Instance Details		
Workspace name *	DataMsc	
Location *	West Europe	
Pricing Tier * ③	Trial (Premium 14-Days Free DBUs)	
Review + Create Next :	Networking >	
<		د

Figure 9: Databricks Set-up

6. If happy with details, click Create

Azure Databricks Service - Mic X	+			- ø ×
↔ ↔ ↔ ↔	🛛 🖨 https://portal	azure.com/?quidstart=true#create/Microsoft.Databricks		··· 🖂 🎝 🕪 🗇 🗉
Microsoft Azure		P Search resources, services, and doos (G+/)	 E & Q @	? © x16140818@student.nci @
Home > New >				
Azure Databricks Se	ervice			×
Validation Succeeded				
*Basics Networking Tags	Review + Create			
Summary				
Basics				
Workspace name	DataMsc			
Subscription	Pay-As-You-Go			
Resource group	MscData			
Location	West Europe			
Pricing Tier	trial			
Networking				
Deploy Azure Databricks workspace your own Virtual Network (VNet)	in No			
Create Previous : Tags				

Figure 10: Databricks Set-up

7. If the resource is created successfully you will be brought to below page. Click on Home to return to main page.

\Lambda Microsoft.Databricks   Overvie: X 🕂		- a ×
↔ ở ŵ	A https://portal.azure.com/?quidstart=true#blade;HubsExtension/DeploymentDetailsBlade;/overview;fd/%2Fsubsorptions%2F59b04986-4751-41c3-b28c-18443c899b9%2	FresourceGroup:%2FMscData%2Fproviden:%2F) 🚥 😇 🏠 🖍 🔟 🕸 🔶 🗮
Microsoft Azure	P Search resources, services, and docs 10+/)	📘 🕞 🧬 🐵 ? 🗇 x16140818@student.nci 🌘
terms     terms	s Overview	Control of the second

Figure 11: Databricks Set-up

8. Next a Storage Container resource, JSON files and Key Vault need to be added. Return to Home page and click on Storage accounts

🔥 Home - Microsoft Azure 💦 🗶 Databricks	× +			- a ×
(←) → C <sup>a</sup> @	portal azure.com/?quickstart=true#home			··· © ☆ II\ 🖸 🏶 🅈 🗉
Microsoft Azure	P Search resources, services, and	docs (S+J)	E 6 0 ?	C x16140818@student.nci
	Azure services ++ Create a resource Recent resources Azure Dadarids Al resources Sat	Image: Socriptions         Image:	App Services More services	
	Name	Туре	Last Viewed	
	S DataMsc	Azure Databricks Service	15 minutes ago	
	📍 Pay-As-You-Go	Subscription	40 minutes ago	
	📍 Azure for Students	Subscription	50 minutes ago	

Figure 12: Storage Account Set-up

9. Click Add

▲ Storage accounts - Microsoft A ×	+			- 0	a x
<>> ℃ @	A https://portal.azure.com/#blade/HubsExtension/BrowseResource/resourceType/Microsoft.Storage%2FStorageAccounts	110%	… ⊠ ☆ ⊻	lin 🖸 🏽 🕈 💕 🕻	
Microsoft Azure	/P Search resources, services, and docs (G+/)			x16140818@student.nc NATIONAL COLLEGE OF IRELA	
Home > Storage accounts National College of Ireland + Add  Manage view ~					×
Fiter by name	stoription == (all) Resource group == (all) X Location == (all) X <sup>1</sup> Add filter	No gro	uping	✓ List view	×

Figure 13: Storage Account Set-up

10. Update below information ensuring Subscription, Resource group and Location matches the Databricks resource. When updated click Review + create

🔥 Create storage account - Micro 🗙	Chusters - Databricks × +	- <b>σ</b> ×
(← → ♂ @	https://portal.azure.com/?quickstart=true#create/Microsoft.StorageAccount-ARM	··· 🖾 🕁 👘 🗉 🔿
Microsoft Azure	P. Search resources, services, and docs (G+/)	🖸 😡 🖓 😳 👘 🖓 🔊
Home > New >		
Create storage accourt	nt	×
Basics Networking Advanced	Tags Review + create	
Azure Storage is a Microsoft-managed redundant. Azure Storage includes Azur Azure Tables. The cost of your storage a Learn more about Azure storage account	ervice providing cloud storage that is highly available, secure, durable, scalable, and Elobs (objects), Xaure Data Lake Storage Gen2, Azure Files, Azure Queues, and count depends on the usage and the options you choose below. ds of	
Project details		
Select the subscription to manage deplo all your resources.	yed resources and costs. Use resource groups like folders to organize and manage	
Subscription *	Pay-As-Yoo-Go	
Resource group *	MscData V	
Instance details		
The default deployment model is Resou using the classic deployment model inst	rce Manager, which supports the latest Azure features. You may choose to deploy ad. Choose classic deployment model	
Clarana account same * O	detector	
Location 1	Conservation of Conservation	
Excauon -	Europe) west Europe	
Performance U		
Account kind 💿	StorageV2 (general purpose v2)	
Replication	Locally-redundant storage (LRS)	
Access tier (default) 💿	Cool O Hot	
Review + create	< Previous Next : Networking >	
https://portal.azure.com/?quickstart+true#		6

Figure 14: Storage Account Set-up

#### 11. Click Create

🔥 Create storage account - Micro 🗙	Chutters - Databricks × +			- 0	×
(←) → @ @	🛛 🗞 https://portal.azvre.com//quickstart=true#create/Microsoft.StorageAccount-ARM			- 🖂 🖻 🗇 🕈	=
Microsoft Azure	20-Search resources, services, and doss (6-7).	8 <b>6</b> 0	◎ ?	ATIONAL COLLEGE OF IRELAN	
Home > New >					
Create storage accou	nt				$\times$
Validation passed					
Basics Networking Advance	Tags Review + create				
Basics					
Subscription	Pav-61-You-Go				
Resource group	NscData				
Location	West Europe				
Storage account name	datasetmisc				
Deployment model	Resource manager				
Account land	StorageV2 (general purpose v2)				
Replication	Locally-redundant storage (LRS)				
Performance	Standard				
Access tier (default)	fot				
Networking					
Connectivity method	Public endpoint (all networks)				
Advanced					
Secure transfer required	Enabled				
Blob soft delete	Disabled				
File share soft delete	Disabled				
Blob change feed	Disabled				
Versioning	Disabled				
Hierarchical namespace	Disabled				
NPS V3	Disabled				
Create	evicus Next > Convolute for automation				
					_
					And in case of

Figure 15: Storage Account Set-up

12. When deployment is complete click Go to Resource to add the JSON files to Azure storage.



Figure 16: JSON Files Upload

13. When in the Storage account click on Storage Explorer

🔥 datasetmsc - Microsoft Azure 🛛 🔘 🖉	Dusters - Databricks × +			- 0 >
(<) → ୯ ŵ	🛛 💊 https://portal.azure.com/?quickstart=true#0	@studentndrf.onmicrosoft.com/resource/subi	criptions/59b04986-4751-41c3-b28c-1844	43c899b9l/resourcegroups/MscData/providers/Microsoft.Storage/storageAccounts/dat= 🚥 😇 🏠 🖍 🖾 👁 📌 🗦
Microsoft Azure		Search resources, services, and docs (0+/)		El 💀 🖉 🕲 ? 😳 x16140818@student.nd
Home > Microsoft.StorageAccount-20	200603130051   Overview >			
storage account				×
P Bearch (Chil+/)	🔛 Open in Explorer 🔿 Move 🖒 Refresh	Delete 😳 Feedback		
Overview     Activity log	Classic alerts in Azure Monitor is announced to reti	re in 2021, it is recommended that you upgrade you	classic alert roles to retain alerting functionality with	db the new alerting platform. For more information, are Continue alerting with ARM storage accounts. If
Pa Access control (IAM)	Resource group (change) : MscData	Performance/Access tier 1 Standard/Hot Replication 1 Localle-redundant storage (LRS)		
Diagnose and solve problems Tota transfer	Location : West Europe Subscription (change) : Pay-As-You-Go		Ac	Account kind I StorageV2 (general purpose v2)
<ul> <li>Events</li> <li>Storage Explorer (preview)</li> </ul>	Subscription ID : 59b04986-4751-41c3 Tags (change) : Click here to add tag	-b28c-18443c899b9f		
Settings				
Access keys     Geo-replication     CORS	Containers Scalable, cost-effective storage for unstructured data Learn more	File shares Serverless SMIE file shares Learn more	Tables Tabular data storage Learn moré	Courses Courses to truthe Learn more
Configuration				

Figure 17: JSON Files Upload

1. Right click on Blob Containers to create the 3 Blobs to hold the Business, Review and Users data

🔺 datasetmsc   Storage Explorer    🗙 🛞 Clu	esters - Databricks × +							a ×
← → ♂ ŵ	https://portal.azure.com/?quickstart=	true#@studentncirlonmicrosoft.com/resource/subscriptions/59b04986-4751-41c3-b28c-18443c899b9(/resourcegroups/Msc	7ata/providers/Micro	soft.Storage/st	torageAccour	its/data: •••		• =
Microsoft Azure		${\cal P}$ -Search resources, services, and docs (G+/)		E 8	e 🖉 🖉	? ©	x16140818@studen NATIONAL COLLEGE OF IR	it.nci
Home > Microsoft.StorageAccount-2020 datasetmsc   Storage Storage account 9 Search (Ctrl+/) «	e Explorer (preview) &	© Query + Add ⊘ Eds ©-Select All ◯ Column Options × Delors ∑ Table Statistics ① Refresh						×
Correlev  Attiny log  Attiny log  Attiny log  Access centrol (MAN)  Togators and safve problems  Data transfer  Events  Storage Deplorer (preview)	Craft bio container  Craft bio container  Fortex  For	INATIONERY No data available in table	∧ ROWNEY					

Figure 18: JSON Files Upload

🔥 datasetmsc   Storage Explorer   🗙 🤗 Clusters - Databricks 🛛 🗙 🕂			- <b>5</b> ×
(←) → C <sup>4</sup>	ntart=true#@studentnonl.onmicrosoft.com/resource/subscriptions/59b04986-4751-41c3-b28c-18443c899b9f/resourcegraups/MscDat	ta/providers/Microsoft.Storage/storag	jeAccounts/data 🚥 🖾 🏠 💷 🕾 🔹 🗮
	P. Search resources, services, and docs (G+/)		🞐 🐵 🤉 😳 x16140818@student.nci 🌘
Home > Microsoft.StorageAccount-20200603130051 [Overview > datasetmsc   Storage Explorer (preview) Storage account // Search (0111/)	y2 ☐ Guery + Add 2/100 ©-Select All 100 Column Options × Doine Σ Table Statistics € Refresh		Successfully created storage container 1:16 PX Successfully created storage container 'user'.
Coreniew Activity log Access control (AM)	RATIFICMERY No data available in table	∧ ROWEY	

Figure 19: JSON Files Upload

14. Click on each container and then click upload



Figure 20: JSON Files Upload

15. Select the corresponding JSON file for each container downloaded from the Yelp Open Dataset website. Once the JSON files have been uploaded click on Access Keys



Figure 21: JSON Files Upload

16. Copy Storage account name and Key 1 to notepad or word. When complete return to the homepage

▲ datasetmsc   Access keys - Mic × +				×
↔ ♂ @ @ #	🛔 https://pertal.aure.com/#Betudentnici.com/resource/tubsor/pionu/59604986-4751-41c2-b28c-18443d9969/resourceGroupu/MicData/providers/Microsoft.Storage/storageAccounts/datasetmic/kays 🚥 😇 🧙	¥ II\ 🖸	æ 🔹	=
Microsoft Azure	P Search measurem, and data (6+)	16140818@str	ident.nci	. ()
Prome         Storage accounts >           a         datasetmost         Access: Storage accounts >           a         datasetmost         Access: Storage accounts >           p         storage accounts >         access: Storage accounts >           a         Overview         access: Storage accessing Storage           %         Togs: Dispose and solve problems.	s keys: Use access keys to authenticate your applications when making requests to this Azure storage account. Store your access keys securely - for example, using Azure Key Yoult - and don't share them. We recommend regenerating your access keys securely - for example, using Azure Key Yoult - and don't share them. We recommend regenerating your access keys securely - for example, using Azure Key Yoult - and don't share them. We recommend regenerating your access keys securely - for example, using Azure Key Yoult - and don't share them. We recommend regenerating your access keys securely access the your access keys secure to use the new keys. This action will not interrupt access to disk from your virtual machines.	ys regularly. You	are provi	X rided
<ul> <li>Data transfer</li> <li>Events</li> <li>Storage Explorer (preview)</li> <li>Settings</li> </ul>	key Concernent of the set of the			0

Figure 22: Key Vault and Secret (1 of 17)

17. Next an Azure Key Vault and a secret need to be created. On the homepage click on Create Resource

A He	me - Microsoft Azure	× 😝 Databricks × +											-	· ø ×
¢	→ C* @	https://portal.azure.com/?quickstart=t	rue≢home									©		ی ♦ Ξ
=	Microsoft Azure		P Search resources, services,	and docs (G+/)						<b>21</b> (	ş 🖉 🛛	? ©	x16140818@H	itudent.nci 🌘
		Azure services	Azure All resources Databricks	Subscriptions	HDInsight clusters	(initial control of the second	Storage accounts	Virtual machines	CO App Services	-> More services				

Figure 23: Key Vault and Secret (2 of 17)

18. Search for Key Vault, select Key Vault resource and click create

🔨 New - Microsoft Azure 🛛 🗙 🗧	+		- ø ×
<-> ୯ @	A https://portal.azure.com/#create/hub		
Microsoft Azure		P Search resources; services; and docs (S+/)	x16140818@student.nd. NATIONAL COLLEGE OF IRELAND
New		3	×
Azure Marketplace See all Get started Recently created	Popular Windows Server 2016 Datacenter Quickstarts + tutorials		
Al + Machine Learning Analytics	O Ubuntu Server 18,04 LTS Learn more		
Blockchain Compute	Web App Quickstarts + tutorials		

#### Figure 24: Key Vault and Secret (3 of 17)

waterbace - waterbart strate ×							- V
	s//portal.azure.com/#blade/Microsoft_Azure_Mar	ketplace/MarketplaceOffersBlade/sele	ctedMenuItemId/home/searchQuery	/key vault		🖾	
	P Search r	esources, services, and docs (G+/)			E 6		x16140818@student.nc. NATIONAL COLLEGE OF IRELAN
Marketplace 🖈	× (	Pricing : All Operating System :	All Publisher : All				
Showing All Resu	's						
vice Providers Showing All Result tegories Started Key Vault - Machine Learning Microsoft	ts Key Vault Analytics Microsoft	RoçSistem Azure Key Vault KoçSistem Bilgi ve İletiyim	Anjuna Enterprise Enclave for Vault	Anglave Data Vault	Anjuna Unseal Anjuna	CipherTrust Cloud Key Manager v1.6.3	
vice Providers Showing All Rest tegories Stated Key Vault Machine Learning typics (schain)	Is No Key and Choose the Understand your fay 'built usage mesuph Analysis of Key Yourt togs	KoçSistem Abure Key Vault KoçSistem Bigi ve İletiyin Increase sexulty and cerbisi over keys and gasuretis with the support of Abure Key Vault	O Anjuna Enterprise Enclave for Vault Anjuna Andreppine Enclaves Secures Histolicop Vault Against Insider Threats	Anqlave Data Vault Anglave Key management system	C Anjuna Uniseal Anjuna Serure Hanhisorp Vault uniseal tokens	Cipherfrust Cloud Key Manager v1.6.3 Thates decurity Multi-Osoid Entrystein Key Ufergite Management	**

Figure 25: Key Vault and Secret (4 of 17)

🔨 Key Vault - Microsoft Azure 🛛 🗙 🕂				- a ×
↔ → @ @ @ @	$https://portal a {\it sure.com} ({\it H} blade/Microsoft_Acure_Marketplace/MarketplaceOffersBlade/selectedMenuItemId/home/searchQuery/key vaultion ({\it H} a {\it sure.com}) and ({\it H} a {\it$			··· 🖾 🕁 💷 🖉 🗧
	P Search resources, services, and docs (G+/)	2 6	1 ©	2 😳 x16140818@student.nci 👰
Home > New > Markapiton > Key Vault & Moresoft Key Vault : 0: S Moresoft Content Create	vis for later.			×

Figure 26: Key Vault and Secret (5 of 17)

19. Enter the following information on the Create Key Vault page again ensuring Resource group and Region are the same as storage account. When complete click Review + create

🔥 Create key yault - Microsoft 🚈 🗙 🎼	A Databases × +					d ×
(←) → ♂ @	A https://portal.azure.com/#create/Microsoft.Key/sult			… 🖾 🕁	± 11\ CD	* * =
Microsoft Azure	P Search resources, services, and docs (G+))	1 😡	Q (0)	? 🗢	x16140818@st	udent.nd
Home > New > Key Vault >						
Create key vault						×
Basics Access policy Network	king Tags Review + create					
Azure Key Vault is a cloud service used developers to store security informatio which greatly reduces the chances that backed by Hardware Security Modules Level 2 validated. In addition, key vault complete audit trail for compliance. In	to manage type, server, and conflicters key hall definitions the need for in their code, I also you to centralise the totage of your appointed servers may be laaked. Key haal also allows you to security store servers and key. An orthor, the Yeldin code are referrant demonstration security damated by the code of the orthor of the demonstration of the security damated by the code and the security damated by the security damated by the security damated by the barn more					
Project details						
Select the subscription to manage dep all your resources.	loyed resources and costs. Use resource groups like folders to organize and manage					
Subscription *	Pay-As-Your-Go					
Resource group *	MycData V Create new					
Instance details						
Key vault name * 💿	mscdatakey 🗸					
Region *	West Europe					
Pricing tier * ③	Standard 🗸					
Soft delete	(Truble) Disable					
Retention period (days) * 💿	90					
Purge protection ③	(Enable (Disable)					
Roview + create	Provins Nest : Access policy >					

Figure 27: Key Vault and Secret (6 of 17)

20. When validation has passed click Create

🔥 Create key vault - Microsoft A 🛛 🗙 🌘	Mabineks X +						- 0	i ×
< → ୯ ŵ	D 🔒 https://portal.azure.com/#create/Microsoft.KeyVault				··· 🖾 🕁	± ₩\		• =
Microsoft Azure	P. Search resources, services, and docs (G+2)	(2)	<b>6</b> 0	۲	? 🐵	x16140818 NATIONAL C	S@student. ourse or se	
Home > New > Key Vault >								
Create key vault								×
S Validation passed								
Basics Access policy Networkin	g Tugs. Review + create							
Basics								
Subscription	Pay-As-You-Go							
Resource group	MscData							
Key vault name	mscdatakey							
Region	West Europe							
Pricing tier	Standard							
Enable soft delete	Enabled							
Enable purge protection	Disabled							
Retention period (days)	90 days							
Access policy								
Azure Virtual Machines for deployment	Disabled							
Azure Resource Manager for template	Disabled							
deployment								
Azure Disk Encryption for volume encryption	Disabled							
Azure Disk Encryption for volume encryption	Disabled							
Permission model	Access control list							
Access policies	1							
Networking								
Connectivity method	Public endpoint (all networks)							
Create Create	evicus Next > Deveload a template for automation							3

Figure 28: Key Vault and Secret (7 of 17)

21. When the Key Vault has been created navigate to the overview page and click Secrets

🔥 mscdatakey - Microsoft Azure 🗙 🤒 D	abricks × +		- ø ×
(←) → C @ [	A https://portal.azure.com/#@studentndrl.onmicrosoft.com/resource/subscriptions/59b04986-4751-4	41c3-b28c-18443c899b9f/resourceGroups/MscData/providers/Microsoft.KeyVault/vaults/mscdatakey/overview	… ♡☆ ⊻ \\ 🖸 @ 🗢 Ξ
Microsoft Azure	P Search resources, services, and docs (G+/)		? 😳 x16140818@student.nd 🍘
Home > mscdatakey   Overview > mscdatakey  Key vaut P Search (Ctrl+/) e	B Delete → Move		×
Overview     Activity log	The soft delete feature has been enabled on this key vault. After you soft delete this key vault, it will remain in you this for more details.	ur subscription as a hidden vault. It will get purged after the retention period you specified. You may purge it sconer, or restore the vau	ult, using Azure PowerShell or Azure CLI. Click $\rightarrow$
Pa. Access control (IAM)	Resource group (change) : MscData	DNS Name : https://mscdatakey.vault.azure.net/	í
Tags	Location : West Europe	Sku (Pricing tier) : Standard	
<ul> <li>Diagnose and solve problems</li> <li>Events (preview)</li> </ul>	Subscription (change) : Pay-As-You-Go Subscription ID : 59b04986-4751-41c3-b28c-18443c899b9f	Directory ID : 6edb49c1-b172-4eea-8053-a71d0a25b68c Directory Name : National College of Ireland A	
Settings  Certificates	Monitoring Show data for last: (1 hour 6 hours 12 hours 11 day 7 days 30 days) Click for additional metrics.		

Figure 29: Key Vault and Secret (8 of 17)

22. Click Generate/Import

🔥 mscdatakey   Secrets - Microson X 🔘 Da	atabricks × +				- a ×
(←) → ♂ @	1 🔒 https://portal.azure.com/#@studentno	rl.onmicrosoft.com/resource/subscriptions/59b04986-4751-41	c3+b28c-18443c899b9t/resourceGroups/MscData/providers/M	icrosoft.KeyVault/vaults/mscdatakey/secrets 🛛 😶 🔂	⊻ II\ 🗊 🏶 🕈 🗏
Microsoft Azure		P Search resources, services, and docs (G+/)		D 🗣 🖓 @ ? 😜	x16140818@student.nci
Home > mscdatakey   Overview >					
🔁 mscdatakey   Secret	ts				×
Key vault	Lawrence Dates a	and the second se			
Constant (CUIV)	T Generate/Import	estore Backup			
Activity log	Name	Туре	Status	Expiration Date	
Access control ((AM)	There are no secrets available.				
Tags					
Diagnose and solve problems					
🗲 Events (preview)					
Settings					
📍 Keys					
🚨 Secrets					
🐖 Certificates					

Figure 30: Key Vault and Secret (9 of 17)

23. Update below information. For Value paste Key 1 from Access Keys for Storage account previously copied to notepad. At this point also save the Key name to notepad. Finally click create

→ C <sup>4</sup> ŵ 0 ≜ =0 http:	x = + c:/portal.azure.com/#@studentnoil.onmicrosoft.com/resource/subscriptions/59b04986-4751-41c3-b28c-18443c899	9691/resourceGroups/MscData/providers/Microsoft.KeyVault/vaults/mscdatakey/secrets … 😇 🟠 👱 🛝	
Microsoft Azure	P Search resources, services, and docs (G+)).	E 🕼 🤗 🐵 🤉 😅 x16140818	Østudent.nci
me 🔰 mscdatakey   Overview 🗲 mscdatakey   Secret	3)		
reate a secret			×
Upload options			
Manual	$\sim$		
Name * 🔘			
datastoragekey			
Value * 💿			
Content type (optional)			
Set activation date?			
06/05/2020	3 PM		
(UTC+01:00) Current Time Zone	× 1		
Set expiration date?			
Expiration Date			
06/05/2022	; AM		
(UIC+01:00) Current Time Zone	·		
Enabled? Yes No			

Figure 31: Key Vault and Secret (10 of 17)

24. After the secret has been created navigate to the properties of the Key Vault

🔨 mscdatakey   Secrets - Microso 🗙 🥘	Databricks × +				- 0
< → ୯ @	0 🔒 🗝 https://portal.azure.com/#@	studentncirl.onmicrosoft.com/resource/subscriptions/59b04986-4751-41c3	l-b28c-18443c899b9f/resourceGroups/MscData/providers/	Vicrosoft.KeyVault/vaults/mscdatakey/secrets ***	S 5 7 11 0 8 4
Microsoft Azure		P Search resources, services, and docs (G+J)		📄 🛛 🕼 🖉 🎯 ?	x16140818@student.nci
Home > mscdatakey   Overview >					
mscdatakey   Secret Key vault	ets				×
,P Search (Ctrl+/) «	+ Generate/Import 🚫 Refresh	T Restore Backup			
Overview     Activity log	The secret 'datastoragekey' has b	en successfully created.			
Access control (IAM)	Name	Туре	Status	Expiration Date	
Tags	datastoragekey		✓ Enabled	6/5/2022	
Diagnose and solve problems					
F Events (preview)					
Settings					
🕇 Keys					
Secrets					
Certificates					
Access policies					
do Networking					
III Properties					

Figure 32: Key Vault and Secret (11 of 17)

25. Copy DNS Name & Resource ID to notepad and return to homepage when finished.

🔥 mscdatakey   Properties - Micro X	Databricks × +		- 0	×
↔ ở ŵ	0 🖨 🗝 https://portal.azure.com/#@stu	dentricif.com/resource/subscriptions/S9b04986-4751-41c3-b28c-18443c899b91/resourceGroups/MstData/providers/Microsoft.KeyVault/vaults/mscdatakey/properties 🚥 🦁 🏚	± IN ED @ •	♦ =
Microsoft Azure		🔎 Search resources, services, and docs (G+/) 💿 🚱 😌 📀 ? 😳	x16140818@student.r NATIONAL COLLEGE OF IRE	
Home > mscdatakey   Overview >	erties			×
P Search (Ctrl+/) «	🗟 Save 🗙 Discard 🚫 Refresh			
😨 Overview	Name	mscdatakey		
Activity log	Sku (Pricing tier)	Standard		
Access control (IAM)	Location	westeurope	Copy to clipbo	pard
Diagnose and solve problems	DNS Name	https://mscdatakey.vauit.azure.net/		6
+ Events (preview)	Resource ID	/subscriptions/59b04986-4751-41c3-b28c-18443c899b9f/resourceGroups/MscData/providers/Microsoft/KeyVault/vaults/mscdatakey		Ð
Settings	Subscription ID	59b04986-4751-41c3-b28c-18443c899b9f		8
📍 Keys	Subscription Name	Pay-As-You-Go		D.
Secrets	Directory ID	6edb49c1-bf72-4eea-8b3f-a7fd0a23b68c		ð
Access policies	Directory Name	National College of Ireland		D
do Networking	Soft delete	(Enable Disable)		
III Properties	Retention period (days)	90		
🔒 Locks	Purge protection	(Evable (Disable)		

Figure 33: Key Vault and Secret (12 of 17)

26. On the homepage click on the previously created Databricks resource

A Home - Microsoft Azure X +						- o ×
← → C ŵ	e.				(110%)	···· © ☆ ⊻ IN ① @ ♦ 🖓 🔤 🗏
Microsoft Azure	P Search resources, services,	and docs (U+/)			D 🕼 O	© ? © x16140818@student.nci
Azure services + Crate a recourse Recent resoure	Ature Databridis Ces	Subscriptions Virtual modvines	Help + support Resource groups	All resources HD	insight More services	
Name		Тура		Last	Viewed	
🕐 msodatakey		Key va	ult	46 m	inutes ago	
datasetmsc		Storag	e account	58 m	inutes ago	
📍 Pay-As-You-Go		Subsc	iption	3 ho	urs ago	
🃍 Azure for Studer	nts	Subsc	iption	a mo	nth ago	
😡 DataMsc		Azure	Databricks Service	a mo	nth ago	

Figure 34: Key Vault and Secret (13 of 17)

27. Click Launch Workspace

\Lambda DataNisc - Microsoft Azure 🛛 🗙 🕂		- o x
(←) → C @	https://portal.azure.com/#@studentnorlonmicrosoft.com/resource/subscriptions/59b04986-475	-41c3-b28c-18443c899b9/resourceGroups/MscData/providers/Microsoft.Dstabricks/workspa 🖽 🚥 😨 🏠 👱 🖿 🕲 🔮 👹 🧮
Microsoft Azure	(A) Search resources, services, and dots (G+/)	All All All All All All All All All All
Home >		
Service DataMsc Azure Databricks Service		×
P Bearch (Ctrl 1/) «	Delete	
😡 Overview	Status : Active	Managed Resource Group : databricks-rg-DataMsc:g7ya4mnvmeztu
Activity log	Resource group : MscData	LiRi : https://adb-78413054594211616.azuredatabricks.net
Pg. Access control (IAM)	Location : West Europe	Prizing Tier : premium
Tags	Subscription : Pay-As-You-Go	
a second as	Subscription ID : 59b04986-4751-41c3-b28c-18443c859b9f	
Settings	Tags (change) : Click here to add tags	
Virtual Network Peerings		Α
🛱 Locks		
関 Export template		
Monitoring		
Diagnostic settings		
Support + troubleshooting		
R New support request		Launch Wortspace

Figure 35: Key Vault and Secret (14 of 17)

28. When the workspace has opened copy the URL and open a new tab.

🔥 datasetmsc - Microsoft Azure 🗙 🤤	Databricks × +			- ø ×
< → ୯ ŵ	A https://adb-784130545942116.16.azuredatabricks.net/?o=784130545942	116#		9 🏠 🕪 🖸 🛎 🕈 🗉
Microsoft Azure			PORT	AL x16140818@student ncitlie
Azur Deselocios	😂 Azure Databric	ks	Free trail ends in 13 days. Upgrade t	D Premium in Azure Portal ? Signed in as x16140838@student.ncir User Settings
Po Workspace		Drop files or click to browse		Admin Console Partner Integrations Manage Account Log Out
Recents B Deta	Explore the Quickstart Tutorial Spin up a cluster, run quinies on preliaded data, and display results in 5 minutes.	Import & Explore Data Quickly import data, preview its schema, create a table, and query it in a notebook.	Create a Blank Notebook Create a notebook to start guerying, visualizing, and modeling your data.	Workspaces ✓ DataMsc x16140010@student.ncid
Custers	Common Tasks	Recents	Documentation	

Figure 36: Key Vault and Secret (15 of 17)

29. In the new tab paste the URL copied previously and add #secrets/createScope at the end so that it is in the following format

https://<\location>.azuredatabricks.net/?o=<\orgID>#secrets/createScope

\land Home -	Microsoft Azure 🛛 🗙 🧃	Databricks X 🤤 Databricks X 🕇	- • ×
€⇒	C ŵ	♥ 🔒 https://adb-784130545942116.16.azuredatabricks.net/?o=784130545942116#secrets/createScope	···· 🖾 🕁 🕅 🛈 🏶 🔶 🗏
Micros	soft Azure		PORTAL x16140818@student.ncirl.ie
	HomePage / Create S	ecret Scope	Free trial ends in 12 days. Upgrade to Premium in Azure Portal 📍 🌲
Azure Databricka	Create Secre	t Scope Cancel Create	î
Home	A store for secrets that store type. Learn more	s identified by a name and backed by a specific	
(D) Workspace	Scope Name Ø		
Ø Recents	Manage Principal O		
	Creator	2	
	Azure Key Vault o		
	DNS Name		
Clusters	https://xxx.vault.azure.ni		
	Resource ID		
Jobs	/subscriptions/x000xx0		
120 Modele			

Figure 37: Key Vault and Secret (16 of 17)

30. Update the below information. DNS Name and Resource ID have been previously copied to notepad. When all information has been entered click Create. Azure Databricks has now been successfully configured.



Figure 38: Key Vault and Secret (17 of 17)

# 3.3 Google

1. A Google account can be set up at Account Setup page<sup>4</sup>. By setting up the account the user will have access to products such as Google Colaboratory (Colab), Google Drive (Drive) and Gmail.

G Create your Google Account × +			x x
		The account. All of Google working for you.	
	English (United States) 💌	Help Privacy Terms	

Figure 39: Google Account Set-up

# 3.4 Genesis Cloud

Genesis Cloud is a cloud Graphic Processing Unit (GPU) service that was required to run the models as the candidate's laptop had limited computational resources. An account can be set up via the Genesis Cloud website<sup>5</sup>. Once the account is created a debit or visa card needs to be added to be able to create virtual instance with the required GPU's. After a credit card has been added the user receives \$50 free credit.

4

https://accounts.google.com/signup/v2/webcreateaccount?hl=en&flowName=GlifWebSignIn &flowEntry=SignUp

<sup>&</sup>lt;sup>5</sup> <u>https://www.genesiscloud.com/</u>



Figure 40: Genesis Cloud Account Set-up

# 3.5 Anaconda

Anaconda Individual is an open sourced Python distribution for machine learning that includes applications such as JupyterLab and Spyder. Anaconda can be downloaded from the Anaconda website<sup>6</sup>.

O Individual Edition   Anaconda 🗙	+						- ø ×
< <p>← → ⊂ ⊕</p>	💿 🚔 https://www.anaconda.com/products/indiv	dual					
	O ANACONDA.	Products - Pricing	Solutions - Reso	ources - Blog	Company -	Get Started	
	Q						
	Individual Editio	n					
	Your da	ata sciel	nce				
	toolkit						
	With over 20 million	users worldwide, the c	open-source Individ	ual			
	science and machine	e learning on a single n	nachine. Developed	for			
	solo practitioners, it i thousands of open-s	s the toolkit that equip ource packages and lit	is you to work with braries				
	Download						
	_	_	_				
	2						

Figure 41: Anaconda Download (1 of 3)

<sup>&</sup>lt;sup>6</sup> <u>https://www.anaconda.com/products/individual</u>

1. Scroll down the page and select python 3.7 64-bit Graphical Installer for Windows

O Individual Edition   Anaconda 🗙 👫				- 0
$\leftarrow \rightarrow$ C $\textcircled{0}$	https://www.anaconda.com/products/individual			8 ··· 🛛 🟠 🕪 🕄 🕈
		Anaconda Installer	S	
	Windows 🕊	MacOS 🗯	Linux 🔕	
	Python 3.7	Python 3.7	Python 3.7	
)	64-Bit Graphical Installer (466 MB)	64-Bit Graphical Installer (442)	64-Bit (x86) Installer (522 MB)	
	32-Bit Graphical Installer (423 MB)	64-Bit Command Line Installer (430 MB)	64-Bit (Power8 and Power9) Installer (276 MB)	
	Python 2.7	Python 2.7		
	64-Bit Graphical Installer (413 MB)	64-Bit Graphical Installer (637 MB)	Python 2.7	
	32. Bit Graphical Installer (356 MB)	64-Bit Command Line Installer (409 MB)	64-Bit (x86) Installer (477 MB)	
	oz-on organisa installer (330 mb)	of or commune the instance (403 MD)	64-Bit (Power8 and Power9) Installer (295 MB)	

Figure 42: Anaconda Download (2 of 3)

2. Run the file after download.

· I ☑ ☐ 〒 I Downloads · IIe Home Share View			
→ ↑ ↓ This PC → Local Disk (C:) → Users → N Name → Ouick access	Date modified	Туре	Size
Desktop / TOQAY (1)     Downloads / Anaconda3-2020.02-Windows->	86_64 22/05/2020 11:36	Application	477,450 KB

Figure 43: Anaconda Download (3 of 3)

3. After the file has opened click next



Figure 44: Anaconda Set-up (1 of 21)

4. Click I agree on the Licence Agreement

ANIACONIDA	License Agreement	
ANACONDA.	Please review the license terms before installing Anaconda3 2020.02 (64-bit).	
Press Page Down to see t	ne rest of the agreement.	
End User License Agreem	ent - Anaconda Individual Edition	^
		-
Copyright 2015-2020, An	aconda, Inc.	
All rights reserved under	the 3-clause BSD License:	
This End User License Agr	eement (the "Agreement") is a legal agreement between you	
and Anaconda, Inc. ("Ana (which was formerly know	conda") and governs your use of Anaconda Individual Edition	
If you accept the terms of agreement to install Anacc	the agreement, click I Agree to continue. You must accept the onda3 2020.02 (64-bit).	
aconda, Inc. ———		

Figure 45: Anaconda Set-up (2 of 21)

5. Click Just Me and click next

	Select Installation Type			
ANACONDA.	Please select the type of inst Anaconda3 2020.02 (64-bit).	allation you would	like to perfo	rm for
T				
Install for:				
Just Me (recommended)				
	¥.:			
O All Users (requires adm	in privileges)			
O All Users (requires adm	in privileges)			
O All Users (requires adm	in privileges)			
O All Users (requires adm	in privileges)			
O All Users (requires adm	in privileges)			
O All Users (requires adm	in privileges)			

Figure 46: Anaconda Set-up (3 of 21)

6. Leave default destination folder and click next

O Anaconda3 2020.02 (64-	bit) Setup		—		×
O ANACONDA.	Choose Install Choose the fold	Location er in which to install A	Anaconda3 2	020.02 (64	Hbit).
Setup will install Anaconda folder, click Browse and se	3 2020.02 (64-bit) i lect another folder.	n the following folder Click Next to continu	. To install in .e.	a differen	t
Destination Folder			_		
C: Users New Janacon	da3		Bro	wse	
Space required: 3.0GB					
Space available: 143.9GB					
Anaconda, Inc. ————					
		< <u>B</u> ack	<u>N</u> ext >	Cano	cel

Figure 47: Anaconda Set-up (4 of 21)

7. It is highly recommended to choose Register Anaconda3 as my default Python 3.7. Click Install

Anaconda3 2020.02 (64)	-bit) Setup	10 <u></u>		×
O ANACONDA.	Advanced Installation Options Customize how Anaconda integrates v	with Windows		
Advanced Options				
Add Anaconda3	to my PATH environment variable			
Not recommended. menu and select "Au Anaconda get found cause problems req Register Anacon This will allow other PyCharm, Wing IDE detect Anaconda as	Instead, open Anaconda3 with the Windo naconda (64-bit)". This "add to PATH" opti d before previously installed software, but uiring you to uninstall and reinstall Anacon da3 as my default Python 3.7 programs, such as Python Tools for Visua , PyDev, and MSI binary packages, to aut s the primary Python 3.7 on the system.	ws Start ion makes t may nda. I Studio tomatically		
Anaconda, Inc. ————	< <u>B</u> ack	Install	Can	cel

Figure 48: Anaconda Set-up (5 of 21)

8. Once completed click next

Anaconda3 2020.02 (64-	bit) Setup	-		$\times$
O ANACONDA.	Installation Complete Setup was completed successfully.			
Completed				_
Show <u>d</u> etails				
naconda, Inc				
	< <u>B</u> ack	lext >	Can	cel

Figure 49: Anaconda Set-up (6 of 21)

# 9. And next again

O Anaconda3 2020.02 (64-b	oit) Setup	_		×
O ANACONDA.	Anaconda3 2020.02 (64-bit) Anaconda + JetBrains			
Anaconda and JetBrains environments tightly inte	are working together to bring you An grated in the PyCharm IDE.	aconda-power	ed	
PyCharm for Anaconda i	s available at:			
https://www.anaconda.	com/pycharm			
			PC	
Anaconda, Inc. ————	< <u>B</u> ack	Next >	Can	cel

Figure 50: Anaconda Set-up (7 of 21)

10. And finally click finish. Anaconda is now installed

Anaconda3 2020.02 (64)	-bit) Setup			$\times$
O ANACONDA.	Completing Anacond (64-bit) Setup Thank you for installing Anaconda Here are some helpful tips and res We recommend you bookmark the back to them later.	Individual Ed sources to get se links so yo	D.02 lition. t you start u can refe	ted. r
	< <u>B</u> ack	Einish	Cano	cel

Figure 51: Anaconda Set-up (8 of 21)

11. After installation is complete search for Anaconda Prompt in the search bar.

All Apps Documents Web Mor	e 🔻	R
Best match		
Anaconda Prompt (anaconda3) App		CN_
anaconda_navigator File folder in site-packages Last modified: 22/5/2020, 12:13	>	Anaconda Prompt (anaconda3) <sub>App</sub>
Apps anaconda-project.exe	>	Open G. Bun as administrator
Anaconda Navigator (anaconda3)	>	
<ul> <li>Anaconda Powershell Prompt (anaconda3)</li> </ul>	>	- Pin to Start
Spyder (anaconda3)	>	TRI Pin to taskbar
📁 Jupyter Notebook (anaconda3)	>	En official
Reset Spyder Settings (anaconda3)	>	
Search the web		
A anaconda - See web results	>	
Folders (8+)		
Documents - This PC (5+)		
🔎 anaconda Prompt (anaconda3)		H: C 🕋 🟦 🥅 💆 😒 🍏 🚾

Figure 52: Anaconda Set-up (9 of 21)

12. Open Anaconda prompt, which is similarly to Command Prompt in windows but powered by the Anaconda Distribution. To check if Python has been successfully installed type python.

Anaconda Prompt (anaconda3) - python	:1 <u>.51</u> 0		×
ase) C:\Users\New>nython		-	
chon 3.7.6 (default, Jan 8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. pe "help", "copyright", "credits" or "license" for more information. >	on win	32	

Figure 53: Anaconda Set-up (10 of 21)

#### 13. To exit type exit()



Figure 54: Anaconda Set-up (11 of 21)

14. Next check the location of the Anaconda file path. The location is needed to be able to add to Systems properties.



Figure 55: Anaconda Set-up (12 of 21)

15. To complete setup the above file paths need to be added to System properties. To open run box press Windows Key + R. When the box opens type sysdm.cpl and click OK

🖅 Run			>
	Type the name of a resource, and Wind	program, folder, docum lows will open it for you.	ent or Internet
<u>O</u> pen:	sysdm.cpl		.~

Figure 56: Anaconda Set-up (13 of 21)

16. Click on Advanced

omputer Name	Hardware	Advanced	System Prote	ction Remote	
Windo on the	ws uses the network.	e following inf	omation to ide	ntify your comput	ter
Computer <u>d</u> escrip	tion:	1			
	F	or example: " omputer".	Kitchen Compu	ter" or "Mary's	
Full computer nar	me: D	ESKTOP-NH	7BU6V		
Workgroup:	W		2		
To rename this c	omouter or	change its dr	main or		
To rename this c workgroup, click	omputer or Change.	change its do	omain or	Change	
		100			

Figure 57: Anaconda Set-up (14 of 21)

17. Then click Environmental Variables

28 10 102 1				1227	
Computer Name	Hardware	Advanced	System Protection	Remote	
You must be log	gged on as a	an Administrat	or to make most of t	nese <mark>changes</mark> .	
Defermente					
Fenomance					
Visual effects	, processor s	scheduling, m	emory usage and vir	tual memory	
				Settings	
User Profiles					
Desktop settin	ngs related to	o your sign-in			
				Settings	
			27		
Start-up and F	Recovery				
System start-u	ip, system fa	ilure and deb	ugging information		
				Settings	
			-		
			-		-
			Environm	ent Variables	

Figure 58: Anaconda Set-up (15 of 21)

18. Click on Path and then click Edit

	Value	
OneDrive	C:\Users\New\OneDrive - National College of Ireland	
OneDriveCommercial	C:\Users\New\OneDrive - National College of Ireland	
OneDriveConsumer	C:\Users\New\OpeDrive	_
Path	C:\Users\New\AppData\Local\Microsoft\WindowsApps;C:\Progra	
TEMP	C:\Users\New\AppData\Local\Temp	
TMP	C:\Users\New\AppData\Local\Temp	
	New Edit Delete	
stem variables Variable	Value	
stem variables Variable ComSner	Value	,
stem variables Variable ComSpec DriverData	Value C:\WINDOWS\system32\cmd.exe C:\Windows\System32\Drivers\DriverData	í
stem variables Variable ComSpec DriverData NUMBER OF PROCESSORS	Value C:\WINDOWS\system32\cmd.exe C:\Windows\System32\Drivers\DriverData 4	,
stem variables Variable ComSpec DriverData NUMBER_OF_PROCESSORS OS	Value C:\WINDOWS\system32\cmd.exe C:\Windows\System32\Drivers\DriverData 4 Windows_NT	`
stem variables Variable ComSpec DriverData NUMBER_OF_PROCESSORS OS Path	Value C:\WINDOWS\system32\cmd.exe C:\Windows\System32\Drivers\DriverData 4 Windows_NT C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wb	`
stem variables Variable ComSpec DriverData NUMBER_OF_PROCESSORS OS Path PATHEXT	Value C:\WINDOWS\system32\cmd.exe C:\Windows\System32\Drivers\DriverData 4 Windows_NT C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wb .COM;.EXE;.BAT;.CMD;.VBS;.VBE;JS;JSE;.WSF;.WSH;.MSC	,
stem variables Variable ComSpec DriverData NUMBER_OF_PROCESSORS OS Path PATHEXT PROCESSOR_ARCHITECTURE	Value C:\WINDOWS\system32\cmd.exe C:\Windows\System32\Drivers\DriverData 4 Windows_NT C:\WINDOWS\system32;C:\WINDOWS;C:\WINDOWS\System32\Wb .COM;.EXE;.BAT;.CMD;.VBS;.VBE;JS;JSE;.WSF;.WSH;.MSC AMD64	,

Figure 59: Anaconda Set-up (16 of 21)

#### 19. Click New

%USERPROFILE%\AppData\Local\Microsoft\WindowsApps	New
C:\Program Files\Intel\WiFi\bin\	
C:\Program Files\Common Files\Intel\WirelessCommon\	Edit
C:\Users\New\AppData\Local\Programs\Microsoft VS Code\bin	
C:\Program Files\Oracle\VirtualBox	Browse
	Delete
	Move <u>U</u> p
	Move D <u>o</u> wr
	Edit <u>t</u> ext

Figure 60: Anaconda Set-up (17 of 21)

20. Then add the locations of files that were returned from Anaconda Prompt where conda. When done click ok and ok for the remaining screens.

%USERPROFILE%\AppData\Local\Microsoft\WindowsApps	New
C:\Program Files\Intel\WiFi\bin\	
C:\Program Files\Common Files\Intel\WirelessCommon\	Edit
C:\Users\New\AppData\Local\Programs\Microsoft VS Code\bi	n
C:\Program Files\Oracle\VirtualBox	Browse
C:\Users\New\anaconda3\Scripts	
C:\Users\New\anaconda3\Library\bin	Delete
C:\Users\New\anaconda3\condabin	
	Move Up
	Move D <u>o</u> wr
	Edit <u>t</u> ext

Figure 61: Anaconda Set-up (18 of 21)

21. 20. After above has been completed to check if installation is correct press Windows Key + R again. Enter cmd and click ok.

B Run				×
19	Type the resource	name of a prog	gram, folder, docur will open it for you	ment or Internet
Open:	cmd			~
⊴pen:	cind			~
	1	OK	Count	Province
		UK	Cancel	Browse

Figure 62: Anaconda Set-up (19 of 21)

22. When Command Prompt opens type conda and the second highlighted box should appear if installed correctly.



Figure 63: Anaconda Set-up (20 of 21)

23. To check which version, enter conda -version



Figure 64: Anaconda Set-up (21 of 21)

# 3.6 Weights & Biases

Weights and Biases (W&B) is a developer tool for tracking experiments for deep learning and users can sign up to W&B with their Google account from the applications website<sup>7</sup>



Figure 65: Weights and Biases Set-up

<sup>&</sup>lt;sup>7</sup> <u>https://app.wandb.ai/</u>

## 3.7 Putty

PuTTY is a Secure Shell (SSH) software that is utilised for the secure connection between an SSH Client and an SSH Server. For the research project PuTTY was used to tunnel or forward Jupyter Lab from the candidate's laptop to the Genesis Cloud Instance. The software can be downloaded from the PuTTY website<sup>8</sup>.

1. Click here

📲 Genesis Cloud 🛛 🔍 🗙	Compute Service - Genesis C 🖉 🗙	How to Create 35H Keys With TIX Download PuTTY - a free 55H and IIX +	E	0	) ×
↔ ở ŵ	https://www.putty.org	···· © ☆ 🛝 🗊	8 🕈	8	
		Download PuTTY PuTTY is an SSH and telnet client, developed originally by Simon Tatham for the Windows platform. PuTTY is open source software that is available with source code and is developed and supported by a group of volunteers. You can download PuTTY here.			
		Below suggestions are independent of the authors of PuTTY. They are not to be seen as endorsements by the PuTTY project.			

Figure 66: PuTTY Set-up (1 of 7)

2. Click on below link for 64-bit



Figure 67: PuTTY Set-up (2 of 7)

3. Click save and then run the downloaded file



Figure 68: PuTTY Set-up (3 of 7)

<sup>&</sup>lt;sup>8</sup> <u>https://www.putty.org/</u>

#### 4. Click Next

PuTTY release 0.74 (64-bit)	Setup —	· [		$\times$
	Welcome to the PuTTY releas (64-bit) Setup Wizard	se 0.7	4	
	The Setup Wizard will install PuTTY release your computer. Click Next to continue or C Setup Wizard.	0.74 (64 ancel to a	Hbit) o exit th	ie Ie
<b>D</b>				
	Back Next		Cance	:I

Figure 69: PuTTY Set-up (4 of 7)

5. And click Next again

PuTTY release 0.74 (64-bit) Setup —		$\times$
Destination Folder		
Click Next to install to the default folder or click Change to choose another.	1	<u>e</u>
Install PuTTY release 0.74 (64-bit) to:		
C:\Program Files\PuTTY\		
Change		
<u>B</u> ack <u>Next</u>	Car	ncel

Figure 70: PuTTY Set-up (5 of 7)

6. Click Install

👷 PuTTY release 0.74 (64-bit) Setup		$\times$
Product Features		
Select the way you want features to be installed.	 6	
Install PuTTY files Add shortcut to PuTTY on the Desktop Put install directory on the PATH for command prompts Associate .PPK files with PuTTYgen and Pageant		
This feature requires 0KB on your hard drive. Back State	Can	icel

Figure 71: PuTTY Set-up (6 of 7)

7. Finally, click Finish. Putty is installed successfully

😥 PuTTY release 0.74 (64-bit) Setup		-		$\times$
	Completed the PuTTY r (64-bit) Setup Wizard	elease 0	.74	
	Click the Finish button to exit the S	etup Wizard	-	
	Back Fi	nish	Canc	el

Figure 72: PuTTY Set-up (7 of 7)

# 3.8 Cliget

Cliget is a Firefox extension that enables the downloading of protected files from sources such as Drive to virtual or remote machines. The add-on adds curl, wget or aria2 to Firefox's download dialog so that user can copy the curl command and run on another machine. The extension can be added to Firefox from the browser add-ons page<sup>9</sup>.

1. Click Add to Firefox

T diget-det bis Extension for 1 X + - 0 X				
← → C* @  © A the https://addons.mozilia.org/en-US/itrefor/addon/diget/	E ···· 당 ☆ IIN ID 🏽 🕈 🚾 🚍			
Firefox Browser ADD-ONS Explore <u>Extensions</u> Themes wore~	Extension Workshop & Developer Hub & Teng toor bit Lag and the second seco			
5.99 Units	127 Besiens 4.5 Stare			
cliget       2         by Zaid Abdulla       2         Download login-protected files from the command line using curt, wget or ariaz.       + Add to Friefex         Trob to reincored for accurity through Modiles Recommended Extensions propers. Make sore you toock it before installing.       1         Immore       1				

Figure 73: Cliget Add-On

<sup>&</sup>lt;sup>9</sup> https://addons.mozilla.org/en-US/firefox/addon/cliget/
## 2. Click Add

🔻 cliget - Get this Extension for 🦄	× +							3	- 0	×
<> → ⊂ ŵ	🖸 🚔 📬 🏠 https://addons.mozilla.org/en-U5/firefox/addon/	cliget/				E	🖾 🕁	IN 10 4	) 🕈 🛙	<b>2</b> =
	Add cliget? It requires your permission to:			Exter	nsion Workshop 💐 Devel	oper Hub 🖉 🛛 Register of Log in				
	Access your cata for all web sites     Learn more about permissions	tensions Themes More Y			Q find add-o	ns $ ightarrow$				
	Add Cancel			í						
	<b>2</b>			4,069 Users	112 Reviews	4.7 Stars				
	cliget			5 <b>*</b>		<u>98</u> <u>9</u>				
	by Zaid Abdulla			2 📩 🖡		<u>+</u> 1				
	Download login-protected files from the comman	d line using curi, wget or aria2.	•	1 *		5				
	This is not monitored for security through Mozilla's Learn more	Recommendec Extensions program. Make sure you trust	it befose installing.							

Figure 74: Cliget Download

# **4** Dataset Creation

The two Datasets were created in Azure Databricks. Launch Databricks as previously demonstrated from the Azure Portal Homepage. First a new cluster needs to be created from the Databricks homepage

1. Click New Cluster

🔥 DataMic - Microsoft Azure 🛛 X 🤤 Databricks 🛛 🗙 🕂			- 0 ×
	s.net/7x=784130545942116		··· 🖂 🖞 🗊 🔹 🔶 🗉
Microsoft Azure			PORTAL x16140818@student.ncirl.ie
Aver Aver	Antabricks Futorial Market and display Market data, and display Ma	Free trait ends in 15 days         Experimentation         Concumentation         Concumentation	Upgrade to Premium in Azure Portal 2

Figure 75: Cluster Set-up (1 of 3)

2. Update below information and click Create Cluster

DataM	hac - Microsoft Azure X 🤤 Create Cluster - Databalicitis X 🕂	- 0 ×
€→	C ŵ 0 ≜ https://adb-7841305453942116.16.azuredatabricks.net/70=784130545542116#cmyste/duster	🖂 🖄 🗂 🗶 🎓 🗏
Micro	psoft Azure	PORTAL x15140818@student.ncirl.ie
	Create Cluster	Free trial ends in 14 days. Upgrade to Premium in Azure Portal 📍 🛔
Azum Databricka	New Cluster Cancel Cancel Count Cluster 20 0-112 0.08 Memory 4 Cons. 15.4 CBU Chinese 14.9 CBM Cons. 15.5 CBU	
Home	Custor Hame	UI   JSON
8		
Workspace		
٥	Pool @	
ALC COMMENTS	None	
	Databricks Runtime Version 🛛 Likam more	
Defa	Paurtime: 6.5 (Scala 2.11, Spark 2.4.5)	
	New This Runtime version supports only Python 3	
	Autopic Options	
2005	ETerminate after 120(\$) minutes of inactivity O	
130 Modele	Worker Type 0         Min Workers         Max Workers           Standard DS3_2         144 08 Henry 4 Cores, 873 080/ v         2         81         6         91	
Q	Driver Type	
Starch	Same as worker 14.0.08 Memory, 4 Coree, 0.75 (26U) 👳	



3. When the Cluster has been approved the State will change to running

🗥 DataMo	c - Microsoft Azure X 😔 Clusters - Databricks X 🕂								
€→	C 🏠 🛛 🕅 https://wdb-7841305459421	16.16.azuredatabricks.net/?o=	7841305459421	16#setting/dusters					··· 🖾 🖓 🔟 🖉 🕈 🗉
Micros	soft Azure							Î	PORTAL x16140818@student ncirl ie
Atore Destroits	Clusters Pools						En	re trial ends in 14 days. Upg	rade to Premium in Azure Portal 🤰 🛔
Rome	+ Create Cluster						All Created by me	Accessible by me	QFilm
B									1 clusters, 0 pinned
Workspace	Name	State	Nodes	Driver	Worker	Runtime	Creator	8	Actions
Recents	Dataset	Running O	2	Standard_DS3_v2	Standard_DS3_v2	6.5 (includes Apache Spark	x16140818@student.ncirl.ie	0	- CE
Defa Custers	▼ Automated Clusters				No clusters found				

Figure 77: Cluster Set-up (3 of 3)

4. Return to Homepage and launch Jupyter Notebook

🕂 DataMsc - Microsoft Azure 🛛 🖌	Databricks × +			- <b>o</b> ×
€ → ୯ ଇ	0 A https://adb-784130545942116.16.azuredatabricks.net/?o=7841305459421	16#		··· 🖂 🎝 🖿 🗇 🔹 🕈 🗉
Microsoft Azure				PORTAL x16140818@student.ncirl.ie
Aver Balances Worksaces Receits Cuis Cuis Cuis	Azure Databrick	Deep film or slick to brown	Free teal ends in 16 days	Upgrade to Premium in Azure Portal 📍 🛓
(m)	Common Tasks	Recents	Documentation	
adat	New Notebook	Recent files appear here as you work	C Documentation	
(28) Modelle	Create Table		C Release Notes	
Q	New Cluster		Getting Started	
Search	New Job			

Figure 78: Jupyter Notebook

The following PySpark code was run to create the two datasets

1. Import libraries



Figure 79: Python Libraries

2. To import external Libraries such as MLflow click Import Library

🔥 EataMcc - Microcoft Azure 🛛 🗙	🤪 Databricks 🛛 🗙 🛨			- a ×
(← → ୯ ଇ	A https://adb-784130545942116.16.azuredatabricks.net/?o=7841305459	942116#	[110%] … 🗟 ☆	lin 🖸 🗶 🕈 🚳 🔳 🔳
Microsoft Azure			PORTAL	z16140818@student.ncirl.ie
Aure Databricks	😂 Azure Databric	ks		? DataMsc 💄
Perme Per Workspace O Resents		Droy files or click to browse	{ }] (+)	
Data A	Explore the Quickstart Tutorial Spin up a causer, run queries on preleaded data, and display results in 6 mnutes:	Import & Explore Data Cuckly import data, preview its schema, create a table, and query if in a notebook.	Create a Blank Notebook Create a notebook to start querying, visualizing, and modeling your data:	
Ciustere	Common Tasks	Recents	Documentation	
Jabb Jabb Models Q Search	New Notebook  Create Table  New Couler  New Job  A New Multion Experiment  M Nove Tubry	Creation	C Documentation C Robuston Notes C Octing Started	

Figure 80: External Libraries (1 of 3)

3. Then libraries from PyPI, Maven and CRAM can be imported



Figure 81: External Libraries (2 of 3)

4. The following external libraries were used

😡 Clusters - Dat	tabricks × +					- 0 ×
← → ♂		redatabricks.net/?o	=784130545942116#/setting/clu	usters/0807-105622-scout336/libraries	130% … 🖂 🏠	IIN 🖸 🏽 🅈 😫 🔤 🗏
Micros	oft Azure				PORTAL	x16140818@student.ncirl.ie
	Clusters / Dataset					? DataMsc 🛔
Azure Databricks	Dataset     Zedit     Start	තී Clone	× Delete			^
Home	Configuration Notebooks (0) Libraries Event	Log Spark	UI Driver Logs Metr	ics Apps Spark Cluster UI - Master -		~
B	%3 Uninstall % Install New					
Workspace	□ Name	Type St	tatus	Source		
0	graphframes.graphframes:0.8.0-spark2.4-s_2.11	Maven -				
Recents	🗆 six	PyPI -				
	MLflow	PyPI -				

Figure 82: External Libraries (3 of 3)

5. Mount files from Azure to Databricks. This was why the Key Vault and secret were necessary

Cmd 2	
1 2 3 4	<pre>dbutils.fs.mount( source = "wasbs://business@datasetmsc.blob.core.windows.net", mount_point = "/mnt/business", extra_configs = {"fs.azure.account.key.datasetmsc.blob.core.windows.net":dbutils.secrets.get(scope = "databricksscope", key = "datastoragekey")})</pre>
Ou1 Com	t[16]: True mmand took 25.67 seconds by x16140818@student.ncirl.ie at 24/06/2020, 10:13:12 on unknown cluster

Figure 83: Mount Azure Files (1 of 3)



Figure 84: Mount Azure Files (2 of 3)



Figure 85: Mount Azure Files (3 of 3)

6. Unmount the files if necessary when all code has been run

```
Cmd 5

      1
      dbutils.fs.unmount("/mnt/business")

      /mnt/business has been unmounted.

      Out[13]: True

      Command took 26.57 seconds --- by x16140818@student.ncirl.ie at 24/06/2020, 10:09:59 on unknown cluster
```

#### Figure 86: Unmount Azure Files (1 of 3)



#### Figure 87: Unmount Azure Files (2 of 3)



#### Figure 88: Unmount Azure Files (3 of 3)

#### 7. Read Business, Review and User JSON files as PySpark DataFrames



Figure 89: Read JSON Files (1 of 3)

Cmd 9



#### Figure 90: Read JSON Files (2 of 3)



## Figure 91: Read JSON Files (3 of 3)

#### 8. Star Distribution

	stars 🔺	count 🔺	
	5	3586460	
2	4	1673404	
3	1	1283897	
ţ	3	842289	
5	2	635072	

Figure 92: Star Distribution

9. Filter rows if they have no Text



Figure 93: Filter Text

### 10. Filter rows if they have no Star

Cma 13	
1 rd = review1.filter(review1.stars.isNotNull())	v - x
▶	
Command took 0.04 seconds by x16140818@student.ncirl.ie at 24/06/2020, 10:21:45 on unknown cluster	

#### Figure 94: Filter Star

## 11. Display Star Count



#### Figure 95: Display Star Count

12. Apply stratified sampling to create Dataset 1. As stratified sampling in PySpark is approximation this needed to be run more than once.



	1	140327
2	2	139934
3	3	139374
4	4	139619
5	5	139521

Figure 97: Stratified Sampling Attempt 1 Star Count

## 13. Attempt 2

Cmd 17	
1 data 0.08	<pre>set1a = rd.sampleBy("stars", fractions={1: 0.108788915309219, 2: 0.220551419569688, 3: 0.166960285230534, 4: 38901064499296, 5: 0.0391697229275646}, seed=434)</pre>
🕨 🗐 dat	taset1a: pyspark.sql.dataframe.DataFrame = [business_id: string, cool: long 7 more fields]
Command to	ook 0.13 seconds by x16140818@student.ncirl.ie at 24/06/2020, 10:46:48 on unknown cluster

## Figure 98: Stratified Sampling Attempt 2

st	ars 🔺	count	-
1		139986	
2		140005	
3		140036	
4 4		139997	
		140047	

Figure 99: Stratified Sampling Attempt 2 Star Count

14. Attempt 3. Below was chosen to proceed with for the research project

#### Figure 100: Stratified Sampling Attempt 3

Cmd 2	20				
1	dis	splay(data	set1	b.groupE	y("s
		-1		4	•
		stars		count	
	1	1		139999	
	2	2		139998	
	3	3		140009	
	4	4		140001	
	5	5		139981	
Sho	owing	all 5 rows.			
		al 🔹 🎽			li
Con	mand	took 1.43 m	inut	es by >	(16140



#### 15. Select Business Fields



Figure 102: Select Business Fields

#### 16. Merge dataset1b & bd1



Figure 103: Merge DataFrames

### 17. Replace unnecessary characters, new lines and carriage returns

```
1 d4 = dataset_one.withColumn("text", regexp_replace(col("text"), "[()&*-,]", ""))
```

Command took 0.14 seconds -- by x16140818@student.ncirl.ie at 24/06/2020, 10:58:07 on unknown cluster

#### Figure 104: Remove Whitespaces (1 of 3)

1 d5 = d4.withColumn("text", regexp\_replace(col("text"), "[\n\r]", " "))

Command took 0.07 seconds -- by x16140818@student.ncirl.ie at 24/06/2020, 10:58:09 on unknown cluster

#### Figure 105: Remove Whitespaces (1 of 3)

#### 18. Check Text format is correct

Cmd 24

Figure 106: (1 of 3)

#### 19. Write Dataset 1 to Databricks File Store (DBFS)



Figure 107: Write to DBFS

Next the code for creating Dataset 2 is described. First was to generate the Social Network, Content Informativeness and Review Rating features to be used as variables for Logistic Regression (LR). After selecting the features LR is applied to predict review useful score. GraphFrame was used to calculate the Social Network Features of outDegree and PageRank. It requires two inputs Vertices and Edges. To create the Vertices and Edges for GraphFrame using the User Json file the below code was run.

Cmd 28
<pre>ver = user.select("user_id", "name")</pre>
▶
Command took 0.06 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:02:04 on unknown cluster
Cmd 29
<pre>1 vertices = ver.withColumnRenamed("user_id","id")</pre>
▶
Command took 0.06 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:02:07 on unknown cluster
Cmd 30
<pre>vertices.select("id").count()</pre>
Out[48]: 1968703
Command took 40.53 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:02:14 on unknown cluster

Figure 108: Vertices

1. Create Edges. As the Users friends are in a list, they need to be formatted to have every unique friend on a new row. The columns need to be renamed also to the correct format required by GraphFrames



Figure 109: Edges

Cmd 34	
<pre>1 ed1.select("user_id","friends").show(3, Fal</pre>	se)
+	+
user_id  friends	
+	+
ntlvfPzc8eglqvk92iDIAw oeMvJh94PiGQnx_6GlndPQ	
ntlvfPzc8eglqvk92iDIAw  wm1z1PaJKvHgSDRKfwhfDg	
ntlvfPzc8eglqvk92iDIAw  IkRib6Xs91PPW7pon7VVig	
+	t
only showing top 3 rows	
Command took 0.46 seconds by x16140818@student.ncirl.	ie at 24/06/2020, 11:07:31 on unknown cluster
Cmd 35	
<pre>1 edges = (ed1.withColumnRenamed("user_id","si</pre>	rc").withColumnRenamed("friends","dst"))
	an a fan dha eanna eanna an an an anna ann an anna anna anna anna anna anna anna anna anna anna anna anna anna
edges: pyspark.sql.dataframe.DataFrame = [src: string,	dst: string]
Command took 0.04 seconds by x16140818@student.ncirl.	ie at 24/06/2020, 11:07:45 on unknown cluster

Figure 110: Rename Fields

## 2. Create GraphFrame

Cmd 36	(+)	
1 g1 = GraphFrame(vertices, edges)		▶- v - x
Command took 0.07 seconds by x16140818@student.ncirl.ie	at 24/06/2020, 11:07:49 on unknown cluster	

Figure 111: GraphFrame

## 3. Display outDegree and create DataFrame of id & outDegree

38						
di	splay(gl.outDegrees)					
	id 🔺	outDegree 🔺				
1	wDheoqPISThML4pShhog3g	57				
2	ckDvozHDR5hWgrDRTMYZkQ	135				
3	soG9o5PqmXilKJHpAB777A	279				
4	tg1Eh5J9iqH5Y0ycb1bejw	123				
5	jO44Apni7iJZVVK4HQ60tA	175				
6	4ZfcCa4m5RWvO4EFzfYm1A	3436				
owina	the first 1000 rows					
owing						
₿	.al - 📥 -		4			
mmand	took 1.22 minutes by x1614081	L8@student.ncirl.	ie at 24/06/2020, 11:07:58 on unknown cluste	ar		
39						
ou	td = g1.outDegrees.select("i	id","outDegree'	)			
	outd: pyspark.sql.dataframe.DataFra	me = [id: string, ou	Degree: integer]			
mmand	took 0.06 seconds by x1614081	L8@student.ncirl.	ie at 24/06/2020, 11:09:50 on unknown cluste	er.		

Figure 112: outDegree

4. Run the PageRank algorithm. Create DataFrame of id & PageRank. Merge outDegree and PageRank DataFrames and rename id to user\_id as required by the algorithm.



Figure 113: PageRank

5. Select the necessary fields from User DataFrame. Rename fields to ensure no duplication later. Merge with previous DataFrame

Cmd 44
<pre>1 uuser = user.select("average_stars", "review_count", "useful", "user_id")</pre>
🕨 🗐 uuser: pyspark.sql.dataframe.DataFrame = [average_stars: double, review_count: long 2 more fields]
Command took 0.06 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:42:12 on unknown cluster
Cmd 45
<pre>1 uu = uuser.withColumnRenamed("useful","useful_sent").withColumnRenamed("average_stars","useful_avg_stars")</pre>
Image: pyspark.sql.dataframe.DataFrame = [useful_avg_stars: double, review_count: long 2 more fields]
Command took 0.05 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:42:16 on unknown cluster
Cmd 46
<pre>1 social_network = network.join(uu, on="user_id", how="inner")</pre>
Image: Instant Inst
Command took 0.05 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:42:19 on unknown cluster

Figure 114: Social Network DataFrame

6. Select necessary fields from Business DataFrame. Again, rename fields to prevent naming duplication



Figure 115: Business Average Star Rating

7. Select necessary fields from Review DataFrame. Remove spacing and unnecessary characters from text. Final check to see if text is correctly formatted

Cmd 49
1 ru = review.select("review_id","user_id","business_id","stars","date","text","useful")
Image: string in the string is the string
Command took 0.07 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:42:52 on unknown cluster
Cmd 50
<pre>1 ru2 = ru.withColumn("text", regexp_replace(col("text"), "[\n\r]", " "))</pre>
Image: pyspark.sql.dataframe.DataFrame = [review_id: string, user_id: string 5 more fields]
Command took 0.12 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:42:57 on unknown cluster
Cmd 51
<pre>1 ru3 = ru2.withColumn("text", trim(ru2.text))</pre>
Image: Image:
Command took 0.10 seconds by x161408180student.ncirl.ie at 23/06/2020, 17:53:51 on unknown cluster

Figure 116: Select Review Fields

Cmd 52	
<pre>1 ru3 = ru2.withColumn("text", regexp_replace(col("text"), "[()&amp;*-,]", ""))</pre>	
▶	
Command took 0.06 seconds by x16140018@student.ncirl.ie at 24/06/2020, 11:43:15 on unknown cluster	



Cmd 53
1 ru3.select("text").show(5,False)
As someone who has worked with many museums I was eager to visit this gallery on my most recent trip to Las Vegas. When I saw they would be showing infamous eggs of the House of Faberge from the Virginia Museum of Fine Arts WHFA I knew I had to go! Tucked away near the gelateria and the garden the Gallery is pretty much hidden from view. It's what real estate agents would call "cozy" or "charming" - basically any euphemism for small. That being said you can still see wonderful art at a gallery of any size so why the two s you ask? Let me tell you: pr icing for this while relatively inexpensive for a Las Vegas attraction is completely over the top. For the space and the amount of art you can fit in there it is a bit much. it's not kid the gallery I could not tring them. the security is not trained properly for the show. When the curating and design teams collaborate for exhibitions there is a definite flow. That means visitors should view the art in a certain sequence whether it be by historical period or cultural significance this is how audio guides are usually developed. When I arrived in the gallery I could not tell where to start and security was certainly not helpful. I was told to "just look around" and "do whatever." At such a fine institution I find the lack of know wledge and respect for the art appalling. I an actually horrified this place is still in business. W 3 year old son needed a haircut this past summer and the lure of the 57 kids cut signs got me in the door. We had to wait a few my son noticed ants crawling all over the floor and the furniture. It was disgusting and I should have walked out then. Actually lave turned around and walked out upon entering b ut I didn't. So the older black male stylist finishes the haircut he was doing and it's our turn. I tell him I want a 42 clipper around the back and sides and then hand cut the top into a standard boys cut. Really freaking finishes: the nhaircut he was doing and it's our turn. I tell him I want a 42 clipper around the back and sides and then
s against it. How does this man who has an alleged cosmetology license not know how to use a set of freaking clippers??? I realized almost immediately that he had no idea what he was doin g. No idea at all. After about 10 minutes of watching this guy stumble through it I said "you know what? That's fine." paid and left. All I wanted to do was get out of that scummy joint a nd take my son to a real haircut place. Bottom line: DO NOT GO HERE. RUN THE OTHER WAY!!!!!
I love Deagan's. I do. I really do. The atmosphere is cozy and festive. The shrimp tacos and house fries are my standbys. The fries are sometimes good and sometimes great and the spicy d ipping sauce they come with is to die for. The beer list is amazing and the cocktails are great. The prices are mid-level so it's not a cheap dive you can go to every week but rather a tr





Figure 119: Remove Whitespaces

1 ru5.select("text").show(5,False)	$ \Rightarrow \forall = 1 $
As someone who has worked with many museums I was eager to visit this gallery on my most recent trip to Las Vegas. When I saw they would be House of Faberge from the Virginia Museum of Fine Arts VMFA I knew I had to go! Tucked away near the gelateria and the garden the Gallery is It's what real estate agents would call "cozy" or "charming" - basically any cuphemism for small. That being said you can still see wonderful so why the two s you ask? Let me tell you: pricing for this while relatively inexpensive for a Las Vegas attraction is completely over the tr nt of art you can fit in there it is a bit much. it's not kid friendly at all. Seriously don't bring them. the security is not trained proper ating and design teams collaborate for exhibitions there is a definite flow. That means visitors should view the art in a certain sequence wi lod or cultural significance this is how audio guides are usually developed. When I arrived in the gallery I could not tell where to start an helpful. I was told to "just look around" and "do whatever." At such a fine institution I find the lack of knowledge and respect for the art	showing infamous eggs of the pretty much hidden from view. L art at a gallery of any size op. For the space and the amou rly for the show. When the cur hether it be by historical per dd security was certainly not appalling.
[] am actually horrified this place is still in business. My 3 year old son needed a haircut this past summer and the lure of the \$7 kids cur had to wait a few minutes as both stylists were working on people. The decor in this place is total garbage. It is so tacky. The sofa they hus sofa with giant holes in it. And my son noticed ants crawling all over the floor and the furniture. It was disgusting and I should have walk have turned around and walked out upon entering but I didn't. So the older black male stylist finishes the haircut he was doing and it's our clipper around the back and sides and then hand cut the top into a standard boys cut. Really freaking simple right? NKONG! Rather than use ti lly cut the hair he went down. Using it moving downward doesn't cut hair it just rubs against it. How does this man who has an alleged cosme use a set of freaking clippers??? I realized almost immediately that he had no idea what he was doing. No idea at all. After about 10 minute: through it I said "you know what? That's fine." paid and left. All I wanted to do was get out of that scummy joint and take my son to a real 0 NOT GO HERE. RUN THE CHATENER. THE STATE AND A state the state to do was get out of that scummy joint and take my son to a real	t signs got me in the door. We ad at the time was a pleather ed out then. Actually I should turn. I tell him I want a #2 he clippers and go up to actua tology license not know how to s of watching this guy stumble haircut place. Bottom line: [
	and and and another and a second

Figure 120: Recheck Text

8. Calculate Word Count, Sentence Count and Price Count. Check to see if Price Count is correct

Cmd 57
<pre>1 ru9 = ru5.withColumn('wordCount', size(split(col('text'), ' ')))</pre>
▶ 🕅 ru9: pyspark.sql.dataframe.DataFrame = [review_ld: string, user_ld: string 6 more fields]
Command took 0.07 seconds by x161408180student.ncirl.ie at 24/06/2020, 11:47:47 on unknown cluster
Cmd 58
<pre>1 ru10 = ru9.withColumn('sentenceCount', size(split(col("text"),r"[\.!?]")) - 1)</pre>
▶ 🗐 ru10: pyspark.sql.dataframe.DataFrame = [review_id: string, user_id: string 7 more fields]
Command took 0.06 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:47:51 on unknown cluster
Cmd 59
<pre>1 rull = rul0.withColumn('priceCount', size(split(col("text"),r"\\$")) - 1)</pre>
Image: Instant Inst
Command took 0.07 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:47:54 on unknown cluster

Figure 121: Word, Sentence, Price Count

Cmd 60
1 rull.select("text","priceCount").show(3, False)
helpful. I was told to "just look around" and "do whatever." At such a fine institution I find the lack of knowledge and respect for the art appalling.  0    I am actually horrified this place is still in business. My 3 year old son needed a haircut this past summer and the lure of the \$7 kids cut signs got me in the door. We had to wait a few minutes as both stylists were working on people. The decor in this place is total garbage. It is so tacky. The sofa they had at the time was a pleather sofa with giant holes in it. And my son noticed ants crawling all over the floor and the furniture. It was disgusting and I should have walked out then. Actually I should have turned around and walked out upon entering but I didn't. So the older black male stylist finishes the haircut he was doing and it's our turn. I tell him I want a #2 clipper around the back and sides and then hand cut the top into a standard boys cut. Really freaking simple right? WRONG! Rather than use the clippers and go up to actua lly cut the hair he went down. Using it moving downward doesn't cut hair it just rubs against it. How does this man who has an alleged cosmetology license not know how to use a set of freaking clippers??! I realized almost immediately that he had no idea what he was doing. No idea at all. After about 10 minutes of watching this guy stumble through it I said "you know what? That's fine." paid and left. All I wanted to do was get out of that scummy joint and take my son to a real haircut place. Bottom line: D O NOT GO HERE. RUN THE OTHER WAY!!!!! 1     I love Desagan's. I do. I really do. The atmosphere is cozy and festive. The shrimp tacos and house fries are my standbys. The fries are sometimes good and sometimes great t and the spicy dipping sauce they come with is to die for. The beer list is amazing and the cocktails are great. The prices are mid-level so it's not a cheap dive you can n go to every week but rather a treat when you do. Try it out. You won't be disappointed!  0
Command took 0.49 seconds by x16140818@student.ncirl.ie at 24/06/2020, 11:48:00 on unknown cluster

#### Figure 122: Check Price Count

## 9. Merge rull and buser DataFrames.



#### Figure 123: Merge DataFrames

#### 10. Calculate average review rating.

 Cmd 62

 1
 review\_avg = ru12.groupBy("business\_id").agg(avg("stars"))

 Image: model in the imag

### Figure 124: Review Average Count

#### 11. Merge ru12 with average review rating

social\_data = ru12.join(review\_avg, on="business\_id", how="inner")

Ele social\_data: pyspark.sql.dataframe.DataFrame = [business\_id: string, review\_id: string ... 10 more fields]

Command took 0.09 seconds -- by x16140818@student.ncirl.ie at 24/06/2020, 11:50:46 on unknown cluster

#### Figure 125: Merge DataFrames

▶- v - x

▶• **v** = x

### 12. Merge social data with social network DataFrames

usefulness\_dataset = social\_data.join(social\_network, on="user\_id", how="inner")

Im usefulness\_dataset: pyspark.sql.dataframe.DataFrame = [user\_id: string, business\_id: string ... 15 more fields]

mmand took 8.05 seconds -- by x16140818@student.ncirl.ie at 24/06/2020, 11:50:49 on unknown cluster

#### Figure 126: Merge DataFrame

## 13. Remove timestamp from date field



Figure 127: Remove Field

14. Calculate deviation to average business rating, deviation to average user rating and rename avg(stars) field to review\_avg\_star.



#### Figure 128: Deviation to Average Review Star Rating

#### 15. Calculate Question Count and Exclamation Count



#### Figure 129: Question and Exclamation Count

#### 16. Drop unnecessary fields



Figure 130: Remove Field

17. Convert the year field from string to timestamp and display date range of reviews



Figure 131: Convert Field

18. Calculate average sentence count and drop more unnecessary fields. All fields now have been created for feature selection



#### Figure 132: Sentence Count and Remove Field

#### 19. The DataFrame is next filtered for the four-month period



Figure 133: Filter Date

20. A User Defined Function is created to convert reviews with a useful count greater or equal to 5 to 1 and less than 5 to 0. The new label field is converted to an integer

Cod 77
<pre>1 def convert_rating(useful): 2 if useful &gt;=5: 3 return 1 4 else: 5 return 0</pre>
Command took 0.85 seconds by x161408180student.ncirl.ie at 24/06/2020, 11:59:09 on unknown cluster
1 useful_convert = udf(lambda x: convert_rating(x))
Command took 0.04 seconds by x161408180student.ncirl.ie at 24/06/2020, 11:59:12 on unknown cluster
Cmd 79
<pre>useful12 = useful11.withColumn("label", useful_convert("useful"))</pre>
▶

Figure 134: User Defined Function

Cnd 80	
<pre>useful13 = useful12.withColumn('label', useful12["label"].cast(IntegerType()))</pre>	$\models_{\forall} \lor = x$
Image: Useful13: pyspark.sql.dataframe.DataFrame = [review_id: string, useful: long 14 more fields]	
Command took 0.05 seconds by x161408180student.ncirl.ie at 24/06/2020, 11:59:19 on unknown cluster	

Figure 135: Cast to Integer

21. The distribution of the 1 & 0's was counted, and stratified sampling was applied to achieve a DataFrame of 21,361 reviews. Again, as stratified sampling is an approximation it takes 2 calculations to achieve the closet possible outcome to the desired DataFrame size

label 🔺 count 🔺	
1 1 39088	
2 0 1176748	
showing all 2 rows.	
■ <u> </u>	

Figure 136: Star Count

Cmd 83				
1 u:	seful_score	= useful <mark>1</mark> 3	.samp	<pre>leBy("label", fractions={1: 0.546484854686866, 0: 0.0181525696240826}, seed=435)</pre>
	useful score:	nysnark sol d	atafram	la DataFrama - fraviau vid strijini usafili linni - 1/1 mora fialda)
Comman	d took 0.13 s	econds by	x16140	Stadestudent.ncirl.ie at 24/06/2020, 12:18:84 on unknown cluster
Cmd 84				
1 d	isnlav(usefu	l score gr	oupBy	("]abe)") count())
1 0	rsp tay (usert	11_30010.51	oupby	( daec ).com(())
	label	count		
1	1	21325		
2	0	21216		
Ohavia				
Snowin	g all 2 rows.			
	<del>له ح</del> اله.		Å	
Comman	d took 6.13 m	inutes by	x16140	818@student.ncirl.ie at 24/06/2020, 12:18:07 on unknown cluster

Figure 137: Starfield Sampling Attempt 1

Cmd 85											
1	<pre>useful_score1 = useful13.sampleBy("label", fractions={1: 0.54740740825163, 0: 0.0182766327177616}, seed=435)</pre>										
	useful score1:	nysnark sol data	frame DataFrame :	- Treview id: string use	aful: long 14 more field	46]					
Commar	nd took 0.15 set	conds by x16.	140818@student.nc	irl.ie at 24/06/2020	), 12:55:55 on unknown	n cluster					
Cmd 86											
1 d	isnlav(usefu	scorel.grou	nBv("label").c	ount())							
	(usera										
	label	<ul> <li>count</li> </ul>	<b>A</b>								
1	1	21347									
2	0	21358									
Showin	ng all 2 rows.										
_											
				8							
Commar	nd took 5.80 min	nutes by x16	40818@student.nc	irl.ie at 24/06/2020	9, 12:56:03 on unknown	n cluster					

Figure 138: Starfield Sampling Attempt 2

## 22. Drop unnecessary fields



Figure 139: Remove Field

23. Select independent variables and convert to Pandas DataFrame

Cmd 88	(*)	
1 independent = useful_score2.select(usef	l_score2.columns[1:13])	
) 🗐 independent: pyspark.sql.dataframe.DataFrame = [	wordCount: integer, sentenceCount: integer 10 more fields]	
Command took 0.09 seconds by x16140818@student.nc	irl.ie at 24/06/2020, 13:05:58 on unknown cluster	
Cmd 89		
1 independent_data = independent.toPandas(		
Command took 6.29 minutes by x16140818@student.nc	irl.ie at 24/06/2020, 13:06:04 on unknown cluster	

Figure 140: Independent Variables

24. Plot Correlation with Seaborn Heatmap for the independent variables. Sentence Count and Word Count are highly correlated



Figure 141: Correlation Heat Map

## 25. Check correlation between dependent variable and independent variables

Cmd 92	
1 dependent_corr = use	<pre>tul_score2.select(usetul_score2.columns[1:14])</pre>
dependent_corr: pyspark	sql.dataframe.DataFrame = [wordCount: integer, sentenceCount: integer 11 more fields]
Command took 0.09 seconds t	y x16140818@student.ncirl.ie at 24/06/2020, 13:24:27 on unknown cluster
Cmd 93	
1 for i in dependent_co	rr.columns:
2 if not( isinstance	e(dependent_corr.select(i).take(1)[0][0], six.string_types)):
3 print( "Corre	lation to Label for ", i, dependent_corr.stat.corr('label',i))
Correlation to Label for	wordCount 0.37859007261711686
Correlation to Label for	sentenceCount 0.3403632569098586
Correlation to Label for	priceCount 0.15729660179779512
Correlation to Label for	review_avg_star -0.025294574712295716
Correlation to Label for	outDegree 0.3764977664518094
Correlation to Label for	pagerank 0.17735386909378298
Correlation to Label for	review_count 0.28741146539021306
Correlation to Label for	abs_Business_star 0.012987010501318596
Correlation to Label for	abs_user_star -0.10607260657028005
Correlation to Label for	questionCount 0.13423651830201377
Correlation to Label for	exclamationCount 0.09361952973792916
Correlation to Label for	avg_sentence_length 0.09438460600745945
Correlation to Label for	label 1.0
Command took 2.58 hours by	x161488188student.prirl.4e at 24/06/2020. 13:24:38 on unknown fluster
1000 State 1000 State St	

Figure 142: Correlation Dependent Variable

## 26. Check correlation if Sentence Count is dropped



Figure 143: Remove Sentence Count



Figure 144: Correlation Heat Map

27. Sentence Count was dropped, and all fields are checked for missing values

Cmd 97								(+						
1 fina	l_data :	= useful_s	core2.drop("	sentenceCour	it")									
) 🖩 fin	al_data: p	)yspark.sql.da	ataframe.DataFr	ame = [review_ii	d: string, word	Count: inte	ger 11 more field	s]						
Command t	ook 0.04	seconds t	y x16140818@s1	tudent.ncirl.i	e at 24/06/2	920, 20:33	26 on unknown cl	uster						
Cmd 98														
1 fina	l_data.:	sel <mark>e</mark> ct([co	unt(when(isn	nan(c), c)).a	lias(c) fo	r c in f	inal_data.colu	nns]).toPandas	().head()					
Out[306]	:													
reviev	v_id wor	dCount price	eCount review	_ <mark>avg_</mark> star outi	Degree page	erank revi	ew_count_abs_Bu	siness_star abs_	user_star ques	tionCount exclam	ationCount avg_sen	tence_length	abel	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	
														å
Command t	ook 7.01	minutes t	y x16140818@s1	tudent.ncirl.i	e at 24/06/2	920, 22:02	:08 on unknown cl	uster						

Figure 145: Sentence Count and Null Values

28. Combine all independent variables to one column

Cmd 99
<pre>1 numeric_features =     final_data.select("wordCount","priceCount","review_avg_star","outDegree","pagerank","review_count","abs_Business_star","abs_user_star","questionCount","exclamationCount","avg_sentence     _length") 2 3</pre>
Immeric_features: pyspark sql dataframe.DataFrame = [wordCount integer, priceCount integer 9 more fields] Command took 0.09 seconds by x161408188student.nctrl.ie at 24/96/2820, 20:33:51 on unknown cluster
Cmd 100
1 nf = numeric_features.columns
Command took 0.04 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:33:55 on unknown cluster

Figure 146: Combine to Single Column

29. Apply VectorAssembler and StandardScaler to the independent variables

Ced 101
<pre>1 va = VectorAssembler(inputCols=nf, outputCol="features",handleInvalid="skip")</pre>
Command took 0.85 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:34:08 on unknown cluster
Cnd 102
<pre>1 lr1 = va.transform(final_data)</pre>
Ir1: pyspark.sql dataframe.DataFrame = [review_id: string, wordCount: integer 12 more fields]
Command took 0.06 seconds by x161408180student.ncirl.fe at 24/06/2020, 20:34:24 on unknown cluster
Ced 103
<pre>scalerII = StandardScaler(inputCol="features", outputCol="scaledFeatures")</pre>
Command took 0.85 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:34:43 on unknown cluster
Circl 104
<pre>1 lr2 = scalerII.fit(lr1).transform(lr1)</pre>
In 2: pyspark sql dataframe DataFrame = [review_id: string, wordCount: integer 13 more fields]
Command took 19.80 minutes by x16140818@student.ncirl.ie at 24/06/2020, 20:37:04 on unknown cluster

Figure 147: VectorAssembler and StandardScaler

#### 30. Split DataFrame into train and test

Ord 185	
<pre>1 train, test = lr2.randomSplit([0.8, 0.2], seed=346)</pre>	
<ul> <li>Irain: pyspark.sql dalaframe DataFrame = [review_id: string, wordCount integer 13 more fields]</li> <li>Ist: pyspark.sql dalaframe.DataFrame = [review_id: string, wordCount integer 13 more fields]</li> <li>Command took 0.06 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:37:09 on unknown cluster</li> </ul>	

#### Figure 148: Train and Test

31. Apply Random Forest to Train DataFrame and select most important features. From Feature Importance four features make up 91% of the outcome

Cred 106
<pre>1 rf = RandomForestClassifier(labelCol="label", featuresCol="scaledFeatures", numTrees=20)</pre>
Command took 0.05 seconds by x161408180student.ncirl.ie at 24/06/2010, 20:37:11 on unknown cluster
Cnd 187
1 feature_selection = rf.fit(train)
Command took 13.51 minutes by x16140818@student.ncirl.ie at 24/06/2020, 20:37:14 on unknown cluster
Cnd 108
1 feature_selection.featureImportances
Out[270]: SparseVector(11, {0: 0.1814] 1: 0.0099, 2: 0.0065, 3: 0.2817, 4: 0.1066, 5: 0.3996] 6: 0.0004, 7: 0.0003, 8: 0.0118, 9: 0.0007, 10: 0.0012}) Command took 0.05 seconds by x16140818@student.ncirl.fe at 24/06/2020, 20:37:16 on unknown cluster

Figure 149: Random Forrest Feature Importance

32. Select 4 important features. Again, apply VectorAssembler and StandardScaler to original DataFrame of 21,361 reviews



Figure 150: VectorAssembler

Cmd 112
<pre>l lr_selected = va_selected.transform(final_data)</pre>
Image: Selected: pyspark.sql.dataframe.DataFrame = [review_id: string, wordCount: integer 12 more fields] Command took 0.12 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:59:35 on unknown cluster
Cmd 113
<pre>1 lr2_selected = scalerII.fit(lr_selected).transform(lr_selected)</pre>
Image: Instant Inst

Figure 151: StandardScaler

33. Split into Train & Test and apply Logistic Regression (LR). Print Area Under a Receiver Operating Characteristic Curve (AUC). The AUC is the default evaluation metric in PySpark for Binary Classification

Cmd 114
<pre>1 train2, test2 = lr2_selected.randomSplit([0.8, 0.2], seed=348)</pre>
<ul> <li>Image: train2: pyspark.sql.dataframe.DataFrame = [review_id: string, wordCount: integer 13 more fields]</li> <li>Image: test2: pyspark.sql.dataframe.DataFrame = [review_id: string, wordCount: integer 13 more fields]</li> </ul>
Command took 0.05 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:59:50 on unknown cluster
Cmd 115
<pre>1 logistic = LogisticRegression(labelCol="label", featuresCol="scaledFeatures",maxIter=10)</pre>
Command took 0.05 seconds by x16140818@student.ncirl.ie at 24/06/2020, 20:59:52 on unknown cluster
Cmd 116
1 cmodel_lr = logistic.fit(train2)
Command took 6.93 minutes by x16140818@student.ncirl.ie at 24/06/2020, 20:59:55 on unknown cluster

Figure 152: Logistic Regression (1 of 2)



Figure 153: Logistic Regression (2 of 2)



Figure 154: Logistic Regression Results (1 of 5)

34. Calculate Accuracy for LR model



Figure 155: Logistic Regression Results (2 of 5)

35. Calculate True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)



Figure 156: Logistic Regression Results (3 of 5)



Figure 157: Logistic Regression Results (3 of 5)

36. Calculate Recall, Precision and F1 score



Figure 158: Logistic Regression Results (4 of 5)



Figure 159: Logistic Regression Results (5 of 5)

37. Next Cross Validation was implemented to try and improve on the test results. It should be noted that MLflow tracks experiments but even though it was imported and had worked on previous runs didn't work on below attempt. The LR achieved slightly better AUC on training and test so it was decided to continue with the LR model. If the number of folds had been increased better results could have possibly been achieved but due to time restrictions this was not possible.



#### Figure 160: Cross Validation (1 of 2)



#### Figure 161: Cross Validation (2 of 2)



#### Figure 162: Cross Validation Results

38. Next the LR model was applied to the Review DataFrame containing over 8 million reviews.



Figure 163: Logistic Regression Full Review DataFrame (1 of 4)



Figure 164: Logistic Regression Full Review DataFrame (2 of 4)

Cmd 142
1 lru = vall.transform(useful15)
Iru: pyspark.sql.dataframe.DataFrame = [review_id: string, useful: long 15 more fields] Command took 0.14 seconds by x16140818@student.ncirl.ie at 24/06/2020, 21:53:22 on unknown cluster
Cmd 143
<pre>1 lru1 = scalerII.fit(lru).transform(lru)</pre>
▶  ■ Iru1: pyspark.sql.dataframe.DataFrame = [review_id: string, useful: long 16 more fields] Command took 8.08 minutes by x16140818@student.ncirl.ie at 24/06/2020, 21:54:03 on unknown cluster

Figure 165: Logistic Regression Full Review DataFrame (3 of 4)



Figure 166: Logistic Regression Full Review DataFrame (4 of 4)

39. The Dataframe wads filtered for predictions equal to 1 i.e. Useful. The Review ID was selected and merged with Review and Business fields to match format of Dataset 1

Cmd 146 1 dataset\_lr = datset2b.filter(datset2b.prediction == 1) dataset\_Ir: pyspark.sql.dataframe.DataFrame = [review\_id: string, useful: long ... 19 more fields] Command took 0.12 seconds -- by x16140818@student.ncirl.ie at 24/06/2020, 22:36:54 on unknown cluster Cmd 147 1 dataset\_lr1 = dataset\_lr.select("review\_id") dataset\_lr1: pyspark.sql.dataframe.DataFrame = [review\_id: string] Command took 0.05 seconds -- by x16140818@student.ncirl.ie at 24/06/2020, 23:35:40 on unknown cluster

Figure 167: Filter for Useful Reviews



Figure 168: Merge DataFrames

## 40. Distribution of Star Rating

stars 🔺	count 🔺	
5	1015982	
4	836528	
3	467037	
1	411088	
2	278268	

Figure 169: Star Distribution

## 41. Apply Stratified Sampling to create Dataset 2

 Cmd 151

 1
 dataset2a= dataset2II.sampleBy("stars", fractions={1: 0.340559685517456, 2: 0.503112107752239, 3: 0.299762117348304, 4: 0.167358414781095, 5: 0.137797716888685}, seed=436)

 Image: String and took on the second secon

## Figure 170: Stratified sampling Attempt 1 (1 of 2)

	stars 🔺	count 🔺	
1	5	140399	
2	3	140319	
3	2	140047	
1	4	140035	
5	1	139733	

Figure 171: Stratified sampling Attempt 1 (2 of 2)

42. While checking star distribution in Colab it was noticed that the Display function did not calculate the star count correctly and therefore a pandas dataframe was used to see correct output

1 pd6	= datase	t2a <mark>.toP</mark> ar	ndas()																		
▶ (1) Sp	ark Jobs																				
databr	icks/spar	k/python	/pyspar	k/sql/t	types.p	y:1636:	Deprec	cationWa	arning:	Using o	or impor	ting the	ABCs f	from 'co	ollection	s' instead of	from 'colle	ctions.ab	c' is dep	recated,	а
in 3.	8 it will	stop wo	rking																		
recuri	n passener	ild (Tretu	5)																		
mand 1	took 11.63	minutes	. py x101	140818@st	tudent.no	cirl.ie a	at 07/08	/2020, 15	9:35:26 0	n Datase	t										
154																					
pd6	.groupby(	'stars")	.count(	)																	
	0																				
t[138]	]:																				
ut[138	]:	roviow id		data	funnu	tout	unoful	usor id	addrass	aitr	latituda	longitudo	nama	stata	ostogorios	roview count					
t[138 bi	]: usiness_id	review_id	cool	date	funny	text	useful	user_id	address	city	latitude	longitude	name	state	categories	review_count					
it [138 bi tars	]: usiness_id	review_id	cool	date	funny	text	useful	user_id	address	city	latitude	longitude	name	state	categories	review_count					
tars	]: usiness_id 140591	review_id 140591	cool	date	funny 140591	text	useful 140591	user_id	address	city 140591	latitude 140591	longitude 140591	name 140591	state 140591	categories 140480	review_count 140591					
tars 1.0 2.0	]: usiness_id 140591 139406	review_id 140591 139406	cool 140591 139406	date 140591 139406	funny 140591 139406	text 140591 139406	useful 140591 139406	user_id 140591 139406	address 140591 139406	city 140591 139406	latitude 140591 139406	longitude 140591 139406	name 140591 139406	state 140591 139406	categories 140480 139387	review_count 140591 139406					
tars 1.0 2.0 3.0	]: usiness_id 140591 139406 140333	review_id 140591 139406 140333	cool 140591 139406 140333	date 140591 139406 140333	funny 140591 139406 140333	text 140591 139406 140333	useful 140591 139406 140333	user_id 140591 139406 140333	address 140591 139406 140333	city 140591 139406 140333	latitude 140591 139406 140333	longitude 140591 139406 140333	name 140591 139406 140333	state 140591 139406 140333	categories 140480 139387 140319	review_count 140591 139406 140333					
tars 1.0 2.0 3.0 4.0	]: usiness_id 140591 139406 140333 140640	review_id 140591 139406 140333 140640	cool 140591 139406 140333 140640	date 140591 139406 140333 140640	<b>funny</b> 140591 139406 140333 140640	text 140591 139406 140333 140640	useful 140591 139406 140333 140640	user_id 140591 139406 140333 140640	address 140591 139406 140333 140640	city 140591 139406 140333 140640	latitude 140591 139406 140333 140640	longitude 140591 139406 140333 140640	name 140591 139406 140333 140640	state 140591 139406 140333 140640	categories 140480 139387 140319 140630	review_count 140591 139406 140333 140640					
tars 1.0 2.0 3.0 4.0 5.0	]: usiness_id 140591 139406 140333 140640 139943	review_id 140591 139406 140333 140640 139943	cool 140591 139406 140333 140640 139943	date 140591 139406 140333 140640 139943	<b>funny</b> 140591 139406 140333 140640 139943	text 140591 139406 140333 140640 139943	useful 140591 139406 140333 140640 139943	user_id 140591 139406 140333 140640 139943	address 140591 139406 140333 140640 139943	city 140591 139406 140333 140640 139943	latitude 140591 139406 140333 140640 139943	longitude 140591 139406 140333 140640 139943	name 140591 139406 140333 140640 139943	state 140591 139406 140333 140640 139943	categories 140480 139387 140319 140630 139935	review_count 140591 139406 140333 140640 139943					

Figure 172: Pandas Groupby

Cmd 155			
1 d	ataset2b = dat ed=436)	taset2II.sampl	eBy("stars", fractions={1: 0.339128080548853, 2: 0.505255836085344, 3: 0.299050803651049, 4: 0.166596829275834, 5: 0.137853843096231},
) 🗐	dataset2b: pyspa I took 0.04 secon	rk.sql.dataframe.D nds by x161408	HaleFrame = [business_id: string, review_id: string 15 more fields] s180studert.ncfrl.1e at 07/08/2020, 20:04:51 on Dataset
Cmd 156			
1 di	splay(dataset2	2b.groupBy( <mark>"</mark> st	cars").count().sort(desc("count")))
♦ (1) \$	Spark Jobs		
	stars 🔺	count 🔶	
1	2	140652	
2	5	140465	
3	3	139959	
4	4	139414	
5	1	139113	
Showing	all 5 rows		
Command	an J Tows. .all ▼ ▲ I took 9.47 minut	tes by x161408	8180student.nc1rl.1e at 07/08/2020, 17:57:09 on Dataset

Figure 173: Stratified sampling Attempt 2 (1 of 2)

Cmd 157																
1 pd	4 = datase	t2b.toPa	ndas()													
▶ (1) S	park Jobs															
Command	took 11.60	minutes	- by x16.	140818@s	tudent.n	cirl.ie	at 07/08	8/2020, 1	9:35:41 c	n Datase	t					
md 158																
1 nd	4.grouphy(	"stars")	.count(	)												
		, ,		,												
Out[14	1]:															
	business id	review id	cool	date	funny	text	useful	user id	address	city	latitude	longitude	name	state	categories	review count
stars																
1.0	139975	139975	139975	139975	139975	139975	139975	139975	139975	139975	139975	139975	139975	139975	139864	139975
2.0	139979	139979	139979	139979	139979	139979	139979	139979	139979	139979	139979	139979	139979	139979	139960	139979
3.0	139985	139985	139985	139985	139985	139985	139985	139985	139985	139985	139985	139985	139985	139985	139971	139985
4.0	140039	140039	140039	140039	140039	140039	140039	140039	140039	140039	140039	140039	140039	140039	140029	140039
5.0	139989	139989	139989	139989	139989	139989	139989	139989	139989	139989	139989	139989	139989	139989	139981	139989
Command	took 0.60 s	econds	by x1614	40818@st	udent.nc	irl.ie a	t 07/08,	2020, 19	:35:42 on	Dataset						

Figure 174: Stratified sampling Attempt 2 (2 of 2)

Cmd 159																	
1 da se	taset2c = ed=436)	dataset2	II.samp	leBy("	stars",	fracti	ons={1	0.3391	18864995	063, 2	0.5053	331635830	719, 3:	0.2990	9828482419	932, <mark>4: 0</mark> .166	6550433083761, 5: 0.137864675320721};
> 🗐	dataset2c: py	/spark.sql.c	lataframe	DataFra	ame = [bu	siness_ic	l: string, r	eview_id:	string 1	15 more f	ields]						
md 160			-,				, ,	,			-						
1 pd	7 = <mark>d</mark> atase	t2c.toPa	ndas()														
▶ (1) S	park Jobs																
Command	took 12.88	minutes	by x16	140818@s	tudent.n	cirl.ie	at 07/08	8/2020, 2	0:05:04 c	in Datase	et						
Cmd 161																	
1 pd	7.groupby(	"stars")	.count(	)													
Out[14	6]: business_id	review id	cool	date	funny	text	useful	user_id	address	city	latitude	longitude	name	state	categories	review_count	
stars																	
1.0	140001	140001	140001	140001	140001	140001	140001	140001	140001	140001	140001	140001	140001	140001	139890	140001	
2.0	140011	140011	140011	140011	140011	140011	140011	140011	140011	140011	140011	140011	140011	140011	139992	140011	
3.0	140007	140007	140007	140007	140007	140007	140007	140007	140007	140007	140007	140007	140007	140007	139993	140007	
4.0	139987	139987	139987	139987	139987	139987	139987	139987	130087	130087	120097	130087	120097	130087	139977	130087	
4.0									100001	155507	139907	155501	139907	133307	100011	155501	

Figure 175: Stratified sampling Attempt 3

43. Clean the text to transfer to candidate's laptop

Cmd 162
<pre>1 d2 = dataset2c.withColumn("text", regexp_replace(col("text"), "[()&amp;*-,]", ""))</pre>
III d2: pyspark.sql.dataframe.DataFrame = [business_id: string, review_id: string 15 more fields]
Command took 0.10 seconds by x161408180student.ncirl.ie at 07/08/2020, 20:18:21 on Dataset
Cmd 163
<pre>1 d3 = d2.withColumn("text", regexp_replace(col("text"), "[\n\r]", " "))</pre>
III d3: pyspark.sql dataframe.DataFrame = [business_id: string, review_id.string 15 more fields]
Command took 0.04 seconds by x16140818@student.ncirl.ie at 07/08/2020, 20:18:31 on Dataset

## Figure 176: Clean Text

164	
d3.select("text").show(5,False)	
(1) Spark Jobs	
rext	
·	
vent there for dinner. Honestly I would not go back again. The food was good nothing special but the prices kinda high and the portions small. You have to order at leas eafood meat veggies and apps to come close to feeling full. And don't forget the beer, And that's only for 2 people.	t 3-4 items from
iheir work is good quality for moderate prices. Niki the owner is the best in terms of workmanship. The store however is generally very messy and the staff are rude and . They often quote you one price and charge more because they "quoted wrong" in the first place. Communication is also very challenging since everyone doesn't speak gre	not accommodatin at English or Can
ness Chinese so it would be hard to get your exact ideas accredited if you don't speak Mandarin Chinese. Appointment times usually run late as well so expect to start does at 715 é you beled 6120. Super a sale of a latence they give functioner contained here sit.	getting your nail
uone at rish in you booked olise. Expecting approxy for latenessthey all adout customer service there is a service the provide the solution of the solution	. Absolutely the
est crepes in Vegas and in my opinion the best I have had in the US. Do yourself a favor and visit this place soon.	
If you are vegetarian or vegan Fern is an excellent choice for a good place to eat. My wife is vegan so I've eaten at Fern several times with her. Even though I'm not	vegetarian I've#

Figure 177: Cleaned Text

### 44. Write data to DBFS

Figure 178: Write to Databricks File Store

45. To Transfer files from DBFS to local computer the following steps were taken. Click Data



Figure 179: Connect to Data in DBFS

46. Add Data

A DataMec-Microsoft Aure X 😝 creation - DataBricks X 🕇 — 🗆 X											
$\overleftarrow{\bullet} \rightarrow 0$	C 🙆 🛛 🖉 🔒 https	III III III III III III III III III II									
Micros	soft Azure						PORI	L >16140	818@student.nci1.ie		
	Data	Add Data						O	? DataMsc 🛔		
Azure Dalatricks	Databases 🗸 🗸	Tables	andard • Permissions	© Run Al	E	Sched 🗎 Sched	ule Second Comments	Runs	C Revision history		
Home Home Workspace Recents	• You need to create a cluster to access tables	• You need to create a cluster to access tables	22.27, 93, 94, 35 OH UMROWN								

Figure 180: Click Add Data

47. Click DBFS, FileStore, datasets, dataset\_one\_csv and part-000. Copy the highlighted path for part-000. This needs to be done for Dataset 2 also.

🔥 DataMic	• Microsoft Azure 🛛 X 🧐 Create	New Table - Databroks × +				- o	×
(÷) → I	C @	110% ···· 🖓 🖞 🛍 🛄	Ξ				
Micro	soft Azure					PORTAL 116140618@student.ncm	e
Azure Datatericos	Create New Table	ter Data Sources				? DataMsc	
Home B Workspace	Select a file from DBFS @ FileStore Filestore C data	C datasets C import-stage C jars	D d1 csv D d2 csv	<ul> <li>_SUCCESS</li> <li>_committed_677946804314</li> <li>_started_677946804314274</li> </ul>			
Recants	innt	E piots E tables		part 00003 fic 6770468043.			
Clusters							
acol Acol Mosols	¢						
٩	/FileStore <sup>r</sup> datasets/d1.csv/part-	-00000-1/1-6779468043142743973-06568	740-31ad-4acf-8135-75b0ca3ed3ac-24	84-1-c000.csv			

Figure 181: Dataset 1 File Path

48. The following URL was then used to start the download for Dataset 1 10

49. The following URL was then used to start the download for Dataset 2 11

The 2 Datasets have been created and next they were transferred to Colab for further processing

# **5** Data Pre-processing and Exploration Google Colabartory

Google Colab is free cloud-based environment for machine learning using Jupyter Notebook. Colab provides free access to Tensor Processing Unit (TPU) and GPU. A limit of 12GB of Radom Access Memory (RAM) is provided which unfortunately was not enough for running the complete project. As a result, data pre-processing and data exploration was complete with Colab.

 The first step was to upload the two datasets from local drive to the Google Drive website<sup>12</sup>. Click New, File Upload and select the two recently download files from DBFS

 $<sup>\</sup>frac{^{10}}{6779468043142743973-d656e740-31ad-4acf-8135-75b0ca3ed3ac-2484-1-c000.csv}$ 

<sup>&</sup>lt;sup>11</sup><u>https://westeurope.azuredatabricks.net/files/datasets/d2.csv/part-00000-tid-</u> 662439142231073286-bad7a3a4-8a45-4896-8737-cba2757ef27f-145331-1-c000.csv

<sup>&</sup>lt;sup>12</sup> <u>https://drive.google.com/drive/u/0/my-drive</u>
🙆 Databricks Data 🛛 Google Drive 🗵	+					- 0	×
€)→ ୯ û	A https://drive.google.com/drive/u/0/raideis/10pKAKMIQQs08_U0XX8UIk-Ny7574LJ3A			🖂 🖻 🖬	۵ 1	6	=
🛆 Drive	Q Search in Drive			0	3		A
New New	My Drive > Databricks Data ~				⊞	0	(III)
My Drive	Name 🔨	Owner	Last mocified	File size	_	Î	
Shared with me	Datasef) csv	me	.lim 27, 2020 me	601 MR			0
C Recent	Dataset2.csv	me	Jun 27, 2020 me	851 MB			Ľ
☆ Starred							-
🗐 Trash							3



3. Go the Colab website<sup>13</sup> and click File, New Notebook



Figure 183: Launch Notebook

4. Select Tensor Processing Unit (TPU) by clicking Runtime, Change Runtime Type, select Hardware accelerator as TPU and then click save.



Figure 184: TPU Set-up (1 of 2)

<sup>&</sup>lt;sup>13</sup> <u>https://colab.research.google.com/notebooks/intro.ipynb</u>

🗴 Databricks Data - Google Drive X 📀 Welcome to Colaboratory - Col X 😳 Untilted Lipynb - Colaboratory X 🕂		- o ×
← → C <sup>*</sup> ŵ Ø ▲ <sup>±</sup> https://colab.research.google.com/drive/TCawjDuSAIP3ehD9dZ	64epFeC02a58Zqq	130% ···· 日本 🕈 🖉 🚍
CO CutitledO.ipynb 🌣 File Edit View Insert Runtime Tools Help <u>All changes saved</u>		🗖 Comment 🗮 Share 🌣 🔥
+ Code + Text		Connect 👻 🖌 Editing 🗠
o <b>0</b> 1		_ → ⊕ ■ ✿ ┋ ≯ :
	Notebook settings Hardware accelerator TP To get the most out of Colat, avoid using a TPU unless you need one. Learn more Omit code cell output when saving this notebook CANCEL	

Figure 185: TPU Set-up (2 of 2)

5. Once the TPU has been assigned the following Python code was run to process Dataset 1 further. Install necessary libraries.

	0	1 !pip install transformers
	G	Collecting transformers Downloading https://files.pythonhosted.org/packages/27/3c/91ed8f5c4e7ef3227b4119200fc0ed4b4fd965b1f0172021c25701087825/transformers-3.0.2-py3-none-any.whl
L		Tamiiyamant alyaadu aatiafiadu yamuaata in (yawllaaal)(likilmutkan) (ldiat naakamaa (fyam tyanafaymaya) (2) 2) 0) Figure 186: Install Transformers Library



Figure 187: Import Libraries

6. Mount Google Drive to access Dataset 1. Go to the URL link to copy authentication code



Figure 188: Mount Google Drive

7. Select text and stars, check for null values, convert star column to integer and display star distribution



Figure 189: Star Distribution

8. Reduce star rating by 1, rename stars column labels, split train & test and reset index



Figure 190: Format Labels

9. Split the training dataset into training and validation. Display star distribution for training, validation and testing and rest index



Figure 191: Training Star Distribution







Figure 193: Test Star Distribution

[]	1 d1_t	rain.reset_index(drop=True)	
C∗		text	labels
	0	I was admittedly a bit skeptical about the Duc	3
	1	Came here for lunch on a weekday. Heard mostly	2
	2	This place is weird. It's in a beautiful old b	2

Figure 194: Reset Train Index

[]	1 d1_6	eval.reset_index(drop=True)	
Ŀ		text	labels
	0	Le service quoique sympathique est si lent que	1
	1	We went back for a second time really looking $\ldots$	2
	2	My boyfriend and I visited this buffet on a Tu	3



[]	1 d1_	_test.reset_index(drop=True)	
C∙		text	labels
	0	I really liked the beer and patio seating here	3
	1	I was very excited when I saw the 'Coming Soon	1
	2	This gets a one star because there's nothing I	0

Figure 196: Reset Test Index

#### 10. Write train, validation and test to csv in Google Drive

[] 1 dl\_train.to\_csv('/content/drive/My Drive/Dataset 1 files/dl\_train.csv',index=False, header=True)
[] 1 dl\_eval.to\_csv('/content/drive/My Drive/Dataset 1 files/dl\_eval.csv',index=False, header=True)
[] 1 dl\_test.to\_csv('/content/drive/My Drive/Dataset 1 files/dl\_test.csv',index=False, header=True)

#### Figure 197: Write to Google Drive

#### 11. Import BertTokenizer, select text column of Dataset 1 and reset index

[]	1 from	transformers import BertTokenizer
[]	1 toker	nizer = BertTokenizer.from_pretrained('bert-base-uncased',do_lower_case=True)
C≁	Downloadi	ng: 100% 232k/232k [00:00<00:00, 1.47MB/s]
[]	1 toker	<pre>ns1 = d1[['text']].copy().sample(n=500000)</pre>
[]	1 toker	ns1.reset_index(drop=True)
Ŀ		text
	0	This place is so good!!! They have quite the s
	1	Tried this little French place last night for
	2	It was soft opening the ordering was a disaste

Figure 198: Bert Tokenizer

### 12. Token Distribution

[]	1 input_ids = []
[]	1 lengths = []
[]	<pre>1 for sentences in tokens1.text: 2 encoded_sent = tokenizer.encode( 3 sentences, 4 add_special_tokens=True, 5 ) 6 input_ids.append(encoded_sent) 7 lengths.append(len(encoded_sent))</pre>
₽	Streaming output truncated to the last 5000 lines. Token indices sequence length is longer than the specified maximum sequence length for this model (688 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (544 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (516 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (737 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (516 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (516 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (516 > 512). Running this sequence through the model will result in indexing Token indices sequence length is longer than the specified maximum sequence length for this model (523 > 512). Running this sequence through the model will result in indexing

Figure 199: Token Function

13. Print Min, Max, Mean, Median and percentage of sentences over 512 Tokens and display token distribution

[]	<pre>1 print(' Min Length: {:} tokens'.format(min(lengths))) 2 print(' Max Length: {:,} tokens'.format(max(lengths))) 3 print(' Mean Length: {:,} tokens'.format(np.mean(lengths))) 4 print(' Median Length: {:,} tokens'.format(np.median(lengths)))</pre>
Ŀ	Min Length: 3 tokens Max Length: 3,090 tokens Mean Length: 149.940786 tokens Median Length: 109.0 tokens
[]	<pre>1 number_trunc = sum(i &gt;=512 for i in lengths) 2 3 number_sent= len(lengths) 4 5 percent = float(number_trunc) / float(number_sent) 6 7 print('{:,} of {:,} sentences({:.1%}) in Dataset 1 are longer the 512 tokens.'.format(number_trunc,number_sent,percent))</pre>
C≁	13,065 of 500,000 sentences(2.6%) in Dataset 1 are longer the 512 tokens.

Figure 200: Token Distribution Statistics



Figure 201: Token Distribution

14. The same process as above was applied to Dataset 2

## 6 Fine-grained Yelp Models

The 6 Fine-grained Yelp models were run using JupyterLab and Genesis Cloud. The following steps were taken to set-up an Instance and run the models

1. Create Instance via Genesis Cloud website<sup>14</sup>. Click Create New Instance

<sup>&</sup>lt;sup>14</sup> <u>https://compute.genesiscloud.com/dashboard/instances</u>

tute Service - Genesis Cli Service - Generis Cloud	× +					- 0
୯ ଜ	D 🔒 https://compute.genesi	scloud.com/dashboard/instano	ei			··· 🖂 🕅 🔟 🌚 💆
						Service Status Developers API 📲 🙆 🛩
	Instances					
	Instances History					
mpute	Nickname	Region	Public IPv4	Private IPv4	Status	Accelerator
ances	There are no instances yet					
gos						
work						
unies						
ount 👻						
ens						
ens						

Figure 202: Instance Webpage

2. Required information for Instance. Hostname, 2 GPU's, Preconfigured image of PyTorch 1.5, Password (remember to copy to notepad), Install NVIDIA GPU driver 430.50. When previous has been done click create instance.

mpute Service - Genesis Cl	× +	
→ C' û	A https://compute.genesiscloud.com/dashboard/instances/create	🛛 ¢
		Service Status Deve
	Instances	
	← Create a New Instance	On-demand
		* Contact support to s
	data-msc data-msc	Expected Hourly C Beta phase II (509
e		1x Instance
		GPUs
	Select an Instance	vCPUs Memory
	Туре	Disk
	CPU GPU	Create
	NVIDIA® 1080TI AMD® MI25 CC AMD® RX470 CC	
	Configuration	
	GPUs 0/10 NVIDIA@ 1080TI	
•		
	VCPUs B/40 Intel® Silver up-to 3.00 GHz	
	Memory 24 /120 GiB DDR4-2666	
	Disk 80 /80 GiB SSD	
	Number of instances	
	1 🔄 Available: 1 / 3 GPUs	
	Image	
	Base OS Snapshots Preconfigured	
	TensorFlow 2.2 PyTorch 1.5 TensorFlow 1.15 Fastai 1.0	
	Authentication	
	San key     Recommended and secure.	
	Please note: The password will be shown only once. Copy it and save it securely,	
	Instance Configuration	

Figure 203: Create New Instance (1 of 2)

💄 Account 👻	Instance Configuration	
Profile	<ul> <li>Install NVIDIA GPU driver 430.50</li> <li>Note: the instalation will happen during the first minutes of</li> </ul>	of your instance being active. <u>Learn more</u> .
SSH Keys		
API Tokens	Security Groups	Expand security group settings 🛛 👻
Billing		
Referrals		

Figure 204: Create New Instance (2 of 2)

3. When the Instance is running copy IPv4 without the SSH and open PuTTy. Paste the copied IPv4 under Session Host Name. Click Tunnels. Under Source port enter 888 and Destination localhost:8888. Click Add and then open and then yes.

Compute Service - Genesis Cl	× +						- ø ×
€ → ୯ ŵ	Image: Second secon	ompute.genesiscloud.com/dashboard/instances					* * 89 🖾 🗉
1	Instance	s				Service Status — Developers AP	o o 🕗 -
	Instances Hist	ory					
) 📕 Compute	Nickname	Region	Public IPv4	Private IPv4	Status Acc	elerator	
Instances Images Network	data-mic	Iceland Reydjavik 1	194.61.20.71 × eas attornagin94.81.25.71 (00)	192,168,10,96	ACTIVE 2 NV	NDIA® 1060TI	
Volumes							
💄 Account 👻							
Profile							
SSH Keys							
API Tokens							

Figure 205: Ipv4

Category			Category				
Category: Session Logging Terminal Keyboard Bell Features Window Appearance Behaviour Translation Selection Colours Connection Data Proxy Teinet Bill Data	Basic options for your PuTT Specify the destination you want to co Host Name (or IP address) [ubuntu@194.61.20.71] Connectionrype: O Raw O Teinet O Riogin O Load, save or delete a stored session Saved Sessions Default Settings Project	Y session printect to <u>Port</u> 22 <u>SSH</u> Serjal <u>Load</u> <u>Save</u> <u>D</u> elete	Category: → Features ↑ → Window → Appearance → Behaviour → Translation → Selection → Colours → Colours → Colours → Colours → Proxy → Telnet → Riogin → SSH → Kex → Host keys → Opher → Appearance	Options Pot forwarding Local ports a Remote port Forwarded ports L8888 loca Add new forwar Source port Destination © Local © Auto	a controlling SSH p accept connection s do the same (SS in lhost:8888 ded port: 8888 localhost:8888 O Remote O IPv4	s from other hosts H-2 only) Remov Agd Opynamic IPv§	/e
ta-SSH L-Senal	Close window on exit: Always Never Only of	on clean exit	TTY X11 Tunnels Bugs ↓				

Figure 206: PuTTY Configuration

4. When PuTTy opens paste the password previously copied and hit enter. After the password has been accepted to start a browserless JupyterLab enter jupyter lab –nobrowser. The URL for accessing JupyterLab will now be generated but first open command prompt and enter ssh -CNL localhost:8888:localhost:8888 ubuntu@194.61.20.71. The IP address at the end will change for every new Instance. Enter the password and if correct JupyterLab has now been forwarded from a user local machine to the Genesis Cloud GPU's.



Figure 207: PuTTy Server

- 5. After JupyterLab is connected the data stored in Drive needs to be uploaded using Cliget. There are 6 files in total relating to train, evaluation and test data. Right click on a file. Click Download, when the window appears do not click save. Check the Cliget add-on in the top right-hand corner of the browser window. A number should appear indicating that the add-on has generated the curl command. Repeat for the remaining 5 files. Next go to JupyterLab and click terminal. Paste the curl commands in the terminal. The data will be uploaded from Drive to Jupyter.
- 6. After the data has been transferred the 6 models are run. For the configuration manual the code saved by W&B will be used as it good for reporting. The candidate has included both the W&B python files and the Notebook files run in Jupyter for the project. The code and results are displayed next

### 6.1 Experiment 1 ALBERT-base D1

Since the  $31^{st}$  of July PyTorch has been updated to 1.6. For running the models the candidate used PyTorch 1.5 therefore if run after  $31^{st}$  a user must replace !pip install simpletransformers with !pip install simpletransformers==0.45.5. Only the models used as part of the Demo Video have been updated.

```
[ ]: !pip install transformers
!pip install simpletransformers
!pip install wandb
[]: !git clone https://github.com/NVIDIA/apex
[]: cd apex
[ ]: !pip install -v --no-cache-dir --global-option="--cpp_ext" --global-option="--cuda_ext" ./
[ ]: import logging
                  import pandas as pd
                  from sklearn.metrics import accuracy_score, f1_score
import wandb
                   import
                  from simpletransformers.classification import ClassificationArgs. ClassificationModel
                  wandb.login()
                 logging.basicConfig(level=logging.INFO)
transformers_logger = logging.getLogger("transformers")
transformers_logger.setLevel(logging.WARNING)
                  # Preparing train data
                  reparting train def_csv('/home/ubuntu/d1_train.csv', error_bad_lines=False)
eval_df = pd.read_csv('/home/ubuntu/d1_eval.csv', error_bad_lines=False)
test_df = pd.read_csv('/home/ubuntu/d1_test.csv', error_bad_lines=False)
                 # Args for model
model_args2 = ClassificationArgs()
model_args2.evaluate_during_training = True
model_args2.evaluate_during_training_steps = 1000
model_args2.eval_batch_size = 32
model_args2.eval_batch_size = 64
model_args2.train_batch_size = 64
model_args2.num_train_epochs = 3
model_args2.ovmrite_output_dir = True
                 model_args2.num_train_epochs = 3
model_args2.overwrite_output_dir = True
model_args2.reprocess_input_data = True
model_args2.best_model_dir = '/home/ubuntu/outputs2/best_model'
model_args2.cache_dir = '/home/ubuntu/outputs2/cache_dir'
model_args2.wandb_project = "Msc"
model_args2.ngpu = 2
model_args2.ngpu = 2
model_args2.use_early_stopping = True
model_args2.early_stopping_metric_minimize = False
model_args2.early_stopping_patience = 5
model_args2.lose_stopping_metric_minimize = False
model_args2.early_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metric_stopping_metr
                  model_args2.learning_rate = 4e-5
                  # Create a TransformerModeL
model_a2 = ClassificationModel("albert", "albert-base-v2", use_cuda=True, num_labels=5, args=model_args2)
                         Train the model
                  model_a2.train_model(
                              train df,
                               eval_df;
eval_df=eval_df,
accuracy= accuracy_score)
                    # Evaluate the mode
                  result, model_outputs, wrong_predictions = model_a2.eval_model(test,accuracy=accuracy_score)
                  wandb.log({'AlbertD1_2': result})
[ ]: result, model_outputs, wrong_predictions = model_a2.eval_model(test_df,accuracy=accuracy_score)
[ ]: print(result)
print(model_outputs)
wandb.log({'AlbertD1_2': result})
```

Figure 208: ALBERT-base D1 Code

C JupyterLa	ıb X	ALBERT-base D1   Msc - Weight X					- σ
$\leftrightarrow \rightarrow 0$	C' û	0 🔒 https://app.wandb.ai/brown	er55/Msc/runs/2iq6a0i9/overview?workspace=user-browner55			120%	··· 🖂 🖞 🗖 🖉 🖉
<b>i</b> #! (		Projects $>$ Msc $\triangle$ $>$ Runs $>$					
(1)	ALBERT-ba	se D1 🖉					:
28	What makes this	s run special? 🖉					
يسر	Privacy						
₹ <sub>mm</sub> ⊧	Tags	+					
>	Author	A browner55					
D	State	finished					
	Start time	July 18th, 2020 at 11:25:17 p	om				
{}	Duration Run path	1m 52s					
	Hostname	deep-learning-msc					
	OS	Linux-5.0.0-41-generic-x86	64-with-debian-buster-sid				
	Python version	3.7.6					
	Python executal	ble/usr/local/share/anaconda3	/bin/python				
	Git repository	git clone https://github.co	m/NVIDIA/apex				
	Git state	git checkout -b "ALBERT-bas	e-D1* 3104fd59776a470a5dee3f0d5f3168cf9ab35b53				
	Command	<python file="" main="" no="" with=""></python>	-f /home/ubuntu/.local/share/jupyter/runtime/kernel-	a8ce8503-3	10a-4f2b-8df9-c9310dca6799.json		
	System Hardwa	re GPU type GeForce GTX 1080 T	1				
	W&B CLI Version	0.9.3					
	Config			Raw	Summary		Raw
	Config paramete	are daecriba vaur madal'e inpute	Logra more	now	Summary matrice describe your results. It	arn mara	New
	comy puramete	ers describe your moder's inputs	Learnmore		Summary means describe your results. Le	earn more	
	Q Search		Page 1 of 4 < >		Q Search		Pagelof1 < >
	Name		Value		Name	Value	
	@ adam_eps	silon	1.000e-8	^	all AlbertD1_2.accuracy	0.6749	
	Ø best_mod	el_dir	/home/ubuntu/outputs2/best_model	- 1	AlbertD1_2.eval_loss	0.7661	
	② cache_dir		/home/ubuntu/outputs2/cache_dir		all AlbertD1_2.mcc	0.5939	
	custom_la	oyer_parameters	D		all confusion_matrix	table-file	
	© custom_p	arameter_groups	٥		all pr	table-file	0
	Ido_lower_	_case	false	. 1	all roc	table-file	9
	@ early_stop	oping_consider_epochs	false				
	early_stop	oping_delta	0				
	early_stop	oping_metric	mcc				0
	early_stop	oping_metric_minimize	false	- 1			
	early_stop	oping_patience	5				
	encoding		2				
	eval_batch	h_size	32				
	evaluate_e	during_training	true				
	evaluate_	during_training_silent	true	~			

Figure 209: ALBERT-base D1 Results

### 6.2 Experiment 2 ALBERT-base D2

	import wandb
•	
•	andb.login()
	rue
	umport torch
	git clone https://github.com/NVIDIA/apex
•	d apex
j	pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./
10.00	m <b>port</b> logging im <b>port</b> pandas <b>as</b> pd from sklearn.metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel
1 1	ogging.basicConfig(level=logging.INFO) :ransformers_logger = logging.getLogger("transformers") :ransformers_logger.setLevel(logging.WARNING)
7 1 6 1	<pre># Preparing train data :rain_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) ?val_df1 = pd.read_csv('/home/ubuntu/d2_eval.csv', error_bad_lines=False) rest_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False)</pre>
7 10 10	<pre># Args for model args4 = ClassificationArgs() args4.evaluate during training = True</pre>
	args4. manual seed = 4
-	args4.eval_batch_size =32
	<pre>ings4.max_seq_length = 235 ings4.train_batch_size = 64</pre>
	<pre>irgs4.num_train_epochs = 3 args4.overwrite output dir = True</pre>
é	<pre>irgs4.reprocess_input_data = True</pre>
	irgs4.best_model_air = '/home/ubuntu/outputs/best_model' args4.output dir = '/home/ubuntu/outputs/output'
-	<pre>args4.cache_dir = '/home/ubuntu/outputs/cache_dir' args4.vandh_providet = "Med"</pre>
-	<pre>imgs4.wandb_kwargs = {"name": "AlbertD2"}</pre>
	<pre>ings4.n_gpu = 2 pres4 use early stepping = True</pre>
-	<pre>args4.case_early_stopping_metric = "mcc"</pre>
	<pre>irgs4.early_stopping_metric_minimize = False args4_early_stopping_nationee = 5</pre>
	args4.learning_rate = 4e-5
7	<pre># Create a TransformerModeL model a4 = ClassificationModel("albert" "albert-base-y2" use cuda=True num labels=5 args=args4)</pre>
	# Train the model
r	nodel_a4.train_model(
	<pre>train_dt1, eval_df=eval_df1,</pre>
	accuracy= accuracy_score)
7	<pre>f Evaluate the model result4, model_outputs4, wrong_predictions4 = model_a4.eval_model(test_df1,accuracy=accuracy_score)</pre>
•	<pre>vandb.log({'AlbertD2': result4})</pre>

Figure 210: ALBERT-base D2 Code

C JupyterLa	ab X	Albert-base D2   Msc - Weight: × +					- σ
$\leftrightarrow \rightarrow 0$	C' û	🛛 🔒 https://app.wandb.ai/brown	her55/Msc/runs/3pk2pk3j/overview?workspace=user-browner55			120%	♡ ☆ 비\ ① @ 🔶 💅 🔤
111 C	Q browner55 >	Projects > Msc 合 > Runs >	Albert-base D2 > Overview				
	Albert-bas	e D2 🖉					I
N	What makes thi	s run special? 🖉					
يسر	Privacy						
₹,,,,Ĕ	Tags	+					
>	Author	A browner55					
6	State	finished					
U	Start time	August 13th, 2020 at 9:47:1.	l am				
{ }	Duration	8s					
	Run path	browner55/Msc/3pk2pk3j					
	Hostname	gc-boring-noyce					
	OS Dathannainn	Linux-5.0.0-41-generic-x86	64-with-debian-buster-sid				
	Python version	3.7.6	3/bin/puthon				
	Git repository	git clone https://github.co	m/NVIDIA/apex				
	Git state	git checkout -b "Albert-bas	se-D2* 5d9b5cbc2fb0e0e525c067723d66f1827440c7a8				
	Command	<pre><python file="" main="" no="" with=""></python></pre>	-f /home/ubuntu/.local/share/jupyter/runtime/kerne	1-979e23b0-	7541-47f8-a2dd-6d77df161a3d.json		
	System Hardwa	re GPU type GeForce GTX 1080	Tì				
	W&B CLI Version	n 0.9.4					
	Config paramete	ers describe your model's inputs	Page 1 of 4		Summary metrics describe your results. L	earn more	Page 1 of 1
	Name		Value		Name	Value	
	@ adam .on	silon	1 0000-8	^	al AlbertD2 accuracy	0.6839	0
	best mod	lel dir	/home/ubuntu/outputs/best_model	-	AlbertD2.eval loss	0.7418	<b>U</b>
	🛞 cache dir		/home/ubuntu/outputs/cache_dir		AlbertD2.mcc	0.605	
	© custom la	aver parameters	()	-	confusion matrix	table-file	
	© custom p	parameter groups	0		al pr	table-file	-
	@ dataloade	er num workers	6	-	all roc	table-file	
	Ø do_lower.	_case	false	-			
	early_stop	pping_consider_epochs	false	-			
	early_stop	pping_delta	0				0
	@ early_stop	pping_metric	mcc				
	early_stop	pping_metric_minimize	false				
	early_stop	pping_patience	5				
	encoding		a:				
	eval_batc	h_size	32				
	evaluate_	during_training	true	v			

Figure 211: ALBERT-base D2 Results

# 6.3 Experiment 3 ELECTRA-base D1

[]:	<pre>!pip install transformers !pip install simpletransformers !pip install wandb</pre>
[]:	import wandb
[]:	wandb.login()
	True
[]:	import torch
[]:	<pre>!git clone https://github.com/NVIDIA/apex</pre>
[]:	cd apex
[]:	<pre>!pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./</pre>
[]:	<pre>import logging import pandas as pd from sklearn.metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel</pre>
	<pre>logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger("transformers") transformers_logger.setLevel(logging.WARNING)</pre>
	<pre># Preparing train data train_df = pd.read_csv('/home/ubuntu/d1_train.csv', error_bad_lines=False) eval_df = pd.read_csv('/home/ubuntu/d1_eval.csv', error_bad_lines=False) test_df = pd.read_csv('/home/ubuntu/d1_test.csv', error_bad_lines=False)</pre>
	<pre># Args for model model_args11 = ClassificationArgs()</pre>
	<pre>model_args11.evaluate_during_training = True model_args11.evaluate_during_training_steps = 1000</pre>
	<pre>model_args11.manual_seed = 4 model_args11 eval_batch_size =32</pre>
	<pre>model_args11.eval_vatch_size =&gt;&gt;&gt; model_args11.max_seq_length = 150</pre>
	<pre>model_args11.train_batch_size = 64 model_args11.train_batch_size = 3</pre>
	model_args11.overwrite_output_dir = True
	<pre>model_args11.reprocess_input_data = True model_args11.reprocess_input_data = true</pre>
	<pre>model_args11.best_model_dir = '/home/ubuntu/outputs/best_model model_args11.output_dir = '/home/ubuntu/outputs/output'</pre>
	<pre>model_args11.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args11.rache_nore.t = "Mrc"</pre>
	<pre>model_args11.wandb_kwargs = { "name": "ELECTRA D1"}</pre>
	model_args11.n_gpu = 2
	<pre>model_args11.use_eary_stopping = rue model_args11.eary_stopping metric = "mcc"</pre>
	<pre>model_args11.early_stopping_metric_minize = False</pre>
	<pre>model_args11.leariy_stopping_pattence = 5 model_args11.learing_nate = 40-5</pre>
	* Crasta a TransformerModel
	<pre># create a fransjormermodel model11 = ClassificationModel("electra", "google/electra-base-discriminator", use_cuda=True, num_labels=5,</pre>
	<pre># Train the model model11.train_model(</pre>
	train_df, eval df=eval df
	accuracy=accuracy_score)
	<pre># Evaluate the model result11, model_outputs11, wrong_predictions11 = model11.eval_model(test_df,accuracy=accuracy_score)</pre>
	<pre>wandb.log({'ELECTRA D1': result11})</pre>
	۶

Figure 212: ELECTRA-base D1 Code

C JupyterLab	x x	ELECTRA-base D1   Msc - Weig × +					- σ :
(←) → G	° û	🛛 🔒 https://app.wandb.ai/brown	erSS/Msc/runs/2owSuhuu/overview?workspace=user-brownerSS			120%	···· 🖂 🖒 🗊 😩 🔶 🖾 🔤
iii c		Projects $\rightarrow$ Msc $\triangle$ $\rightarrow$ Runs $\rightarrow$	ELECTRA-base D1 > Overview				
(i)	ELECTRA-b	base D1 🖉					I
N	What makes this	s run special? 🖉					
	Privacy						
3mF	Tags	+					
>	Author	A browner55					
n	State	finished					
6	Start time	July 22nd, 2020 at 3:20:00 p	m				
{ }	Duration	55					
	Run path	browner55/Msc/2ow5uhuu					
	Hostname	deep-learning-msc	a shi dabi sha sa si d				
	US Puthon vorsion	Linux-5.0.0-41-generic-x86_	64-With-debian-buster-sid				
	Python executal	5.1.0	/bin/python				
	Git repository	git clone https://github.co	m/NVIDIA/apex				
	Git state	git checkout -b "ELECTRA-ba	se-D1* 0ac5dd6253d3954a4875b5648a9af439b5ae7b4e				
	Command	<pre>cpython with no main file&gt;</pre>	-f /home/ubuntu/.local/share/jupyter/runtime/kernel	-c477132b-	9af7-458c-a8e7-86a1583a7105.json		
	System Hardwa	re GPU type GeForce GTX 1080 T	ſ				
	W&B CLI Version	0.9.3					
	Q Search	ers describe your model's inputs.	Pagelof4 < >		Summary metrics describe your results. Lea	rn more	Pagelof1 < >
	Name		Value		Name	Value	
	ø adam_eps	silon	1.000e-8	^	all ELECTRA D1.accuracy	0.6792	0
	øbest_mod	el_dir	/home/ubuntu/outputs/best_model	- 1	ELECTRA D1.eval_loss	0.7448	
	cache_dir		/home/ubuntu/outputs/cache_dir		all ELECTRA D1.mcc	0.5992	
	© custom_la	oyer_parameters	0		all confusion_matrix	table-file	
	© custom_p	arameter_groups	0		all pr	table-file	0
	Ø do_lower_	_case	false				9
	early_stop	oping_consider_epochs	false				
	early_stop	oping_delta	0				
	early_stop	oping_metric	mcc				0
	early_stop	oping_metric_minimize	false	-			
	early_stop	oping_patience	5				
	encoding						
	eval_batch	h_size	32				-
	evaluate_e	during_training	true				
	evaluate_o	during_training_silent	true	~			

Figure 213: ELECTRA-base D1 Results

# 6.4 Experiment 4 ELECTRA-base D2

<pre>wandb.login() True import torch igit clone https://github.com/NVIDIA/apex cd apex ipip install -vno-cache-dirglobal-option="cuda_ext" -/ import logging import pands as pd from sklearn.metrics import accuracy_score from sinpletransformers.classification import classificationModel logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger("transformers") transformers_logger = logging.getLogger("transformers") transformers_logger = logging.getLogger("transformers") transformers_logger = logging.getLogger("transformers") transformers_logger.setLevel(logging.MANNO) # Preparing train dat train_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) test_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Arps_for model model_args12_evaluate_during_training = True model_args12_evaluate_during_training = True model_args12_evaluate_during_training = True model_args12.imm_train_potch = 3 model_args12.imm_train_potch = 3 model_args12.imm_train_potch = 3 model_args12.imm_train_potch = 5 model_args12.imm_train_potch = True model_args12.imm_train_gm_trie = "mee" model_args12.imm_train</pre>		import wandb
<pre>andb.login() if ue import torch ig it clone https://github.com/NVIDIA/apex id apex it pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ import logging import pandas as pd from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy interval accuracy interval accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics import acc</pre>		
<pre>rue mport torch git clone https://github.com/NVIDIA/apex id apex pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ mport logging mport pandas as pd from sklaen-metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel orgin_basicConfig(level=logging,INF0) rinsformers_logger = loggins,gstLogger("transformers") ransformers_logger.setLevel(logging,WARNING) / Preparing train data rest_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) val_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) val_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) /* Args for model model_args12 = ClassificationArgs() model_args12.evaluate_during_training = True model_args12.evaluate_during_training = frue model_args12.evaluate_during_training = frue model_args12.evaluate_during_training = frue model_args12.evaluate_during_training = frue model_args12.evaluate_during_training = true model_args12.evaluate_during_training = frue model_args12.evaluate_during</pre>	v	<pre>vandb.login()</pre>
<pre>import torch igit clone https://github.com/NVIDIA/apex id apex td apex td apex td apex td apex td apex tpip install -vno-cache-dirglobal-option="cpp_ext"global-options"cuda_ext" ./ import logging import pandas as pd from sklean.metrics import accuracy_score from sklean.metrics import accuracy, error_bad_lines=False) transformers_logger.stkevel(logging.WANNING) # Args for model model_args12.evaluate_during_training = True model_args12.evaluate_during_training = true model_args12.evaluate_during_training = true model_args12.evaluate_during_training = True model_args12.evandb_droing_metric = "mec" model_args12.evandb_fir = '/home/uburtu/outputs/netw_model' model_args12.evandb_fir = '/home/uburtu/outputs/netw_model' model_args12.evandb_fir = '/home/uburtu/outputs/netw_model' model_args12.evandb_fir = #ccc model_args12.evandb_fir = frue model_args12.evandb_fir = frue model_args1</pre>		l'rue
<pre>git clone https://github.com/NVIDIA/apex d apex pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ mport logging mport pandas as pd rem sklearn.metrics import accuracy_score rem simpletransformers.classification import ClassificationArgs, ClassificationModel orgging.basicConfig(level=logging.INFO) ransformers_logger = logging.getLogger("ransformers") ransformers_logger.setLevel(logging.MARNING) <i>Preparing train data</i> rain_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) val_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) est_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) val_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) est_dfi = pd.read_csv('home/ubuntu/d2_test.csv', error_bad_lines=False) est_dfi = pd.read_csv('home/ubuntu/d2_test.csv', error_bad_lines=False) est_dfi = pd.read_csv('home/ubuntu/d2_test.csv', error_bad_lines=False) est_dfi = pd.read_csv('home/ubuntu/d1test.csv', error_bad_lines=False) est_del_args12.evaluate_dring_training = True est_del_args12.veal_bit = 'home/ubuntu/outputs/bst_model' est_args12.veal_bit = 'home/ubuntu/outputs/bst_model' est_args12.veal_bit = 'home/ubuntu/outputs/bst_model' est_args12.veal_bit = 'home/ubuntu/outputs/bst_model' est_args12.veal_bit = 'home/ubuntu/outputs/bst_model' est_args12.veal_bit = 'home/ubuntu/outputs/bst_model' est_args12.veal_bit = 'home/ubuntu/est_est_est_est_est_est_est_est_est_est_</pre>	i	mport torch
<pre>d apex pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ mport logging mport pands as pd rom sklearn.metrics import accuracy_score rom sklearn.metrics import accuracy_score rom simpletransformers.classification import ClassificationArgs, ClassificationModel ogging.basicConfig(level=logging.INFO) ransformers.logger = logging.etLogger("transformers") ransformers.logger.setLevel(logging.WARNING) ' Preparing train data rain_dfl = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) val_dfl = pd.read_csv('/home/ubuntu/d2_tval.csv', error_bad_lines=False) val_dfl = pd.read_csv('ubuntu/d2_tval.csv', error_bad_lines=False) val_dfl = pd.read_csv('ubuntu/outputs/best_model odel_args12.evaluate_during_training = True odel_args12.ivm_train_patch_size = 64 odel_args12.ivm_train_patch_size = 64 odel_args12.ivm_train_patch = 'home/ubuntu/outputs/cathe_dir' odel_args12.ivm_train_potch = 3 odel_args12.ivm_train_potch = 7 odel_args12.ivm_train_potch = 7 odel_args12.ivm_train_potch = 7 odel_args12.ivm_train_grametric = mcc" odel_args12.ivm_train_grametric = "mcc" odel_args12.ieming_rate = 4e-5 ' Create a TransformerModel ' Train the model odel12.rein_model( rein_dfl, curuasy=accuracy_score) ' Evaluate the model</pre>	1	git clone https://github.com/NVIDIA/apex
<pre>pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ mport logging mport pandas as pd rom sklearn.metrics import accuracy_score rom simpletransformers.classification import ClassificationArgs, ClassificationModel ogging.basicConfig(level=logging.INFO) ransformers_logger = logging.getLogger("transformers") ransformers_logger.setLevel(logging.WARNINO) ' Preparing train data' ' Preparing train data' ' Preparin</pre>	c	d apex
<pre>mport logging mport pandas as pd mport pandas project pandas project pandas pandas project pandas pandas project pandas project pandas project pandas project pandas project pandas project pandas project pandas project pandas project pandas project pandas project pandas project pandas pandas project pa</pre>	1	pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./
<pre>logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger("transformers") transformers_logger.setLevel(logging.WARNING) # Preparing train data train_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) test_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args12 = classificationArgs() model_args12.evaluate_during_training_steps = 1000 model_args12.evaluate_during_training_steps = 1000 model_args12.evaluate_during_traine_e = 64 model_args12.evaluate_during_traine_e = 64 model_args12.output_dir = '/home/ubuntu/outputs/output' model_args12.evaluate_during ''home/ubuntu/outputs/cache_dir' model_args12.evaluate_during_traine_iffer iffer model_args12.evaluate_during_mate: "ELECTRA D2"} model_args12.evalus_topping_metric = "mcc" model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_metric = "mcc" model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_patience = 5 model_args12.evalus_topping_metric_minimize = False model_args12.evalus_topping_metric_minimize = False model_args12.evalus_topping_metric_minimize = False m</pre>	1 1 1 1	import logging import pandas as pd from sklearn.metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel
<pre># Preparing train data rrain_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) exst_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args12 classificationArgs() model_args12 cvaluate_during_training = True model_args12.evaluate_during_training_steps = 1000 model_args12.evaluate_during_training_steps = 1000 model_args12.eval_batch_size = 32 model_args12.eval_batch_size = 64 model_args12.oval_batch_size = 64 model_args12.overwite_output_dir = True model_args12.overwite_output_dir = True model_args12.cvaluatic_dir = '/home/ubuntu/outputs/output' model_args12.cvaluatic_dir = '/home/ubuntu/outputs/output' model_args12.uvandb_more(= "Msc" model_args12.uvandb_kwargs = {"mae": "ELECTRA D2"} model_args12.evaly_stopping_metric = "mcc" model_args12.evaly_stopping_metric = %mcc" model_args12.evaly_stopping_metric = %mce" model_args12.evaly_stopping_metric = %more model_args12.evaly_stopping_terric = %more model_ar</pre>	1	<pre>logging.basicConfig(level=logging.INFO) :ransformers_logger = logging.getLogger("transformers") :ransformers_logger.setLevel(logging.WARNING)</pre>
<pre># Args for model model_args12 = ClassificationArgs() model_args12.evaluate_during_training = True model_args12.evaluate_during_training_steps = 1000 model_args12.eval_batch_size = 32 model_args12.eval_batch_size = 32 model_args12.eval_batch_size = 64 model_args12.num_train_epochs = 3 model_args12.overwrite_output_dir = True model_args12.reprocess_input_data = True model_args12.cost_model_dir = '/home/ubuntu/outputs/best_model' model_args12.cost_model_dir = '/home/ubuntu/outputs/best_model' model_args12.cost_model_dir = '/home/ubuntu/outputs/cache_dir' model_args12.cost_model_dir = '/home/ubuntu/outputs/cache_dir' model_args12.wandb_project = "Msc" model_args12.wandb_project = "Msc" model_args12.use_early_stopping = True model_args12.esely_stopping = True model_args12.esely_stopping_metric = "mcc" model_args12.early_stopping_metric = 5 model_args12.early_stopping_patience = 5 model_args12.learly_stopping_patience = 5 model_args12.learly_stopping_traite = 4e-5 # Create a TransformerModel model12 = ClassificationModel("electra", "google/electra-base-discriminator", use_cuda=True, num_l # Train the model model12.train_model( model12.train_model( model12.train_model( model12.train_model( model12.train_model( model2.train_model2.train_model( model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_model2.train_mo</pre>	# t e t	<pre># Preparing train data xrain_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) xval_df1 = pd.read_csv('/home/ubuntu/d2_eval.csv', error_bad_lines=False) xest_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False)</pre>
<pre>Nodel_args12.indSeq_length = 233 nodel_args12.ind_ind_size = 64 nodel_args12.ind_ind_pochs = 3 nodel_args12.overwrite_output_dir = True nodel_args12.ineaf_ind_ind_pochs nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_ind_poch nodel_args12.ineaf_ind_poch nodel_</pre>	# n n n n	<pre># Args for model model_args12 = ClassificationArgs() model_args12.evaluate_during_training = True model_args12.evaluate_during_training_steps = 1000 model_args12.eval_batch_size = 4 model_args12.eval_batch_size = 32 model_args12.eval_batch_size = 32</pre>
<pre>model_args12.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args12.wandb_project = 'Msc" model_args12.wandb_project = 'Msc" model_args12.n_gpu = 2 model_args12.use_early_stopping = True model_args12.early_stopping_metric = 'mcc" model_args12.early_stopping_metric_minimize = False model_args12.early_stopping_patience = 5 model_args12.learning_rate = 4e-5 # Create a TransformerModel model12 = ClassificationModel("electra", "google/electra-base-discriminator", use_cuda=True, num_1 # Train the model model12.train_model( train_df1, aval_df=eval_df1, accuracy=accuracy_score) # Evaluate the model</pre>	rrrr	<pre>Nodel_args12.train_batch_size = 64 model_args12.train_batch_size = 64 model_args12.num_train_epochs = 3 model_args12.overwrite_output_dir = True model_args12.reprocess_input_data = True model_args12.best_model_dir = '/home/ubuntu/outputs/best_model' model_args12.output_dir = '/home/ubuntu/outputs/outputs'</pre>
<pre>model_args12.early_stopping_patience = 5 model_args12.learning_rate = 4e-5 # Create a TransformerModel model12 = ClassificationModel("electra", "google/electra-base-discriminator", use_cuda=True, num_1 # Train the model model12.train_model( train_df1, eval_df=eval_df1, accuracy_score) # Evaluate the model</pre>		<pre>model_args12.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args12.wandb_project = "Msc" model_args12.wandb_kwargs = {"name": "ELECTRA D2"} model_args12.ngpu = 2 model_args12.use_early_stopping = True model_args12.early_stopping metric = "mcc" model_args12.early stopping metric minimize = False</pre>
<pre># Train the model nodel12.train_model( :rain_df1, .val_df=eval_df1, .ccuracy=accuracy_score) # Evaluate the model</pre>	n n f	model_args12.early_stopping_patience = 5 model_args12.learning_rate = 4e-5 f Create a TransformerModel model12 = ClassificationModel("electra", "google/electra-base-discriminator", use_cuda=True, num_lat
# Evaluate the model		<pre># Train the model model12.train_model( train_f1, eval_df=eval_df1, accuracy accuracy score)</pre>
result12, model outputs12, wrong predictions12 = model12.eval model(test d+1.accuracv=accuracv sco		
wandb.log({'ELECTRA D2': result12})		# Evaluate the model result12. model outputs12. wrong predictions12 = model12.eval model(test df1.accuracv=accuracv score

Figure 214: ELECTRA-base D2 Code

C JupyterLab	ь х	ELECTRA-base D2   Msc - Weig X 🕂					- σ
$\leftrightarrow \rightarrow c$	C û	🛛 🔒 https://app.wandb.ai/brown	er55/Msc/runs/321fjyyo/overview?workspace=user-browner55			(120%)	··· 🖂 🖞 🔟 🖉 🔹 🐨 🖾
<b>i</b> ii (	Q browner55 >	Projects → Msc 台 → Runs →	ELECTRA-base D2 > Overview				
	ELECTRA-b	base D2 🖉					:
~	What makes this	s run special? 🖉					
Ö	Privacy						
m	Tags	+					
>	Author	A browner55					
D	State	finished					
	Start time	August 12th, 2020 at 5:46:56	pm				
{}	Rup nath	browner55/Msc/321fippo					
	Hostname	data-msc					
	05	Linux-5.0.0-41-generic-x86_(	i4-with-debian-buster-sid				
	Python version	3.7.6					
	Python executal	ble/usr/local/share/anaconda3,	/bin/python				
	Git repository	git clone https://github.com	n/NVIDIA/apex				
	Git state	git checkout -b "ELECTRA-bas	e-D2* 5d9b5cbc2fb0e0e525c067723d66f1827440c7a8				
	Command	<python file="" main="" no="" with=""></python>	f /home/ubuntu/.local/share/jupyter/runtime/kerne	-3eac4f75-	5096-43bc-9585-b805f1eaa038.json		
	System Hardwa	re GPU type GeForce GTX 1080 T					
	W&B CLI Version	0.9.4					
	Config			Paw	Summary		Paur
	coning	1.11.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1		now			
	comy paramete	ers describe your model's inputs.	Learn more		summary metrics describe your results. Learn	more	
	Q Search		Pagelof4 < >		Q Search		Page 1 of 1 < >
	Name		Value		Name	Value	
	adam_eps	silon	1.000e-8	^	atd ELECTRA D2.accuracy	0.6822	
	ø best_mod	el_dir	/home/ubuntu/outputs/best_model		all ELECTRA D2.eval_loss	0.8643	
	cache_dir		/home/ubuntu/outputs/cache_dir		all ELECTRA D2.mcc	0.6029	
	© custom_la	yer_parameters	0		all confusion_matrix	table-file	
	Custom_p	arameter_groups	0	_	ail pr	table-file	0
	Ø dataloade	r_num_workers	6		all roc	table-file	9
	Ø do_lower_	_case	false				
	early_stop	oping_consider_epochs	false	_			
	<pre>@ early_stop</pre>	oping_delta	0				0
	early_stop	oping_metric	mcc				
	early_stop	oping_metric_minimize	false				
	early_stop	oping_patience	5				
	encoding		2				0
	<pre>@ eval_batcl</pre>	h_size	32				
	evaluate_	during_training	true	~			

Figure 215: ELECTRA-base D2 Results

### 6.5 Smaller Bert D1

	bib Tupcart Maudo
-	import wandb
۷	<pre>vandb.login()</pre>
1	rue
1	import torch
1	git clone https://github.com/NVIDIA/apex
•	d apex
4	<pre>pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./</pre>
100	import logging
1	import pandas as po from sklearn.metrics import accuracy_score
1	rom simpletransformers.classification import classificationArgs, classificationModel
	<pre>logging.basicConfig(level=logging.INFO)</pre>
1	ransformers_logger = logging.getLogger( transformers ) ransformers_logger.setLevel(logging.WARNING)
7	# Preparing train data
-	:rain_df = pd.read_csv('/home/ubuntu/d1_train.csv', error_bad_lines=False) eval_df = pd.read_csv('/home/ubuntu/d1_eval.csv', error_bad_lines=False)
-	<pre>cest_df = pd.read_csv('/home/ubuntu/d1_test.csv', error_bad_lines=False)</pre>
7	<pre># Args for model prest = ClassificationArgs()</pre>
	args7.evaluate_during_training = True
	rgs7.evaluate_during_training_steps = 1000 rgs7.emanual seed = 4
	urgs7.eval_batch_size =32
	args7.max_seq_length = 150
-	rgs7.num_train_epochs = 3
-	args7.overwrite_output_dir = True
	nggs/.reprocess_input_data = inue nggs/.best model dir = '/home/ubuntu/outputs/best model'
	args7.output_dir = '/home/ubuntu/outputs/output'
	<pre>args7.cache_dir = '/home/ubuntu/outputs/cache_dir' prgs7.wardb.prgsigst = "Mes"</pre>
-	rgs7.wandb_kwargs = { "name": "BertMinD1"}
	ngs7.n_gpu = 2
	<pre>srgs/.use_early_stopping = Irue srgs/.early_stopping metric = "mcc"</pre>
	<pre>args7.early_stopping_metric_minimize = False</pre>
	<pre>rrgs7.early_stopping_patience = 5 rgs7.learning rate = 4e-5</pre>
-	args7.do_lower_case = True
7	t Create a TransformerModel
r	<pre>node1_a/ = classificationMode1("bert", "goog1e/bert_uncased_L-10_H-768_A-12", use_cuda=True, num_l;</pre>
7	# Train the model model a7.train model(
Î	train_df,
	eval_df=eval_df, accuracy= accuracy_score)
7	<pre># Evaluate the model accultzmodel_accurrent_accur</pre>
	esure, moder_outputs, whong_predictions, = moder_a.evar_moder(test_dr,acturacy_acturacy_store)
۲ «	<pre>vandb.log({'BertMinD1': result/})</pre>

Figure 216: Smaller Bert D1 Code

💭 JupyterLa	sb X	Smaller BERT D1   Msc – Weight 🗙 🕂					- 6
$( \leftrightarrow ) \rightarrow 0$	C' û	0 A https://app.wandb.ai/browner5	5/Msc/runs/rmegp432/overview?workspace=user-browner55			(120%)	영 ☆ 🕼 🗈 🔹 👷
	Q browner55 >	Projects $>$ Msc $\triangle$ $>$ Runs $>$ Sn	naller BERT D1 > Overview				
(i)	Smaller BE	RT D1 🖉					:
~	What makes thi:	s run special? 🖉					
Ö	Privacy						
X	Tags	+					
<i>1</i>	Author	finished					
D	State Start time	July 30th, 2020 at 3:49:05 pm					
{}	Duration	7s					
0	Run path	browner55/Msc/rmegp432					
	Hostname	gc-elated-hypatia					
	05	Linux-5.0.0-41-generic-x86_64-	with-debian-buster-sid				
	Python version	3.7.6					
	Python executal	ble/usr/local/share/anaconda3/bi	in/python				
	Git repository	git clone https://github.com/N	WIDIA/apex				
	Git state	git checkout -b "Smaller-BERT-	D1* 459de22d59c64e30fd4b368c368c5b74e269f3dd				
	Command	<pre><python file="" main="" no="" with=""> -f </python></pre>	/home/ubuntu/.local/share/jupyter/runtime/kernel-d	5a5/4/2-3a9c-40c0-	aad5-81ed9//443a1.json		
	System Hardwa	re GPU type GeForce G1X 1080 II					
	Web CEI VEISIOI	0.5.4					
	Config			Raw Surr	nmary		Raw
	Config paramete	ers describe your model's inputs. Le	earn more	Sumr	mary metrics describe your results. Learn more		
	Q Search		Pagelof4 < >		2 Search		Page 1 of 1 < >
	Name		Value	Na	me	Value	
	adam_eps	silon	1.000e-8	ad	BertMinD1.accuracy	0.6719	
	Ø best_mod	lel_dir	/home/ubuntu/outputs/best_model	ad	BertMinD1.eval_loss	0.912	
	cache_dir		/home/ubuntu/outputs/cache_dir	ai	BertMinD1.mcc	0.5901	
	custom_la	ayer_parameters		al	confusion_matrix	table-file	
	custom_p	arameter_groups	0	ai	pr	table-file	
	dataloade	er_num_workers	6	ad	roc	table-file	4
	Ø do_lower_	_case	true				
	early_stop	oping_consider_epochs	false				
	early_stop	oping_delta	0				
	early_stop	oping_metric	mcc				
	early_stop	oping_metric_minimize	false				
	early_stop	oping_patience	5				
	encoding		~				
	<pre>@ eval_batcl</pre>	h_size	32				
	evaluate_	during_training	true	v			

Figure 217: Smaller Bert D1 Results

# 6.6 Experiment 6 Smaller Bert D2

<pre>import wondb iumport wondb iumport wondb iumport wondb iumport torch import torch import torch iumport torch iumport torch iumport logging import pads as pd from sklarn.metrics import accuracy_score.pt.metrics from sklarn.metrics f</pre>		!pip install wandb
<pre>is wandb.login() True True import torch igit clone https://github.com/NVIDIA/apex cd apex igit clone https://github.com/NVIDIA/apex cd apex igit clone https://github.com/NVIDIA/apex cd apex ipip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ ipip install -vno-cache-dirglobal-option="cuda_ext" ./ import logging import promas as pd from silearn.metrics import accuracy_score from simpletransformers.classification import classificationArgs, classificationModel logging.basicConfig(level=logging.INFO) transformers_logger = logging.getUiogger("transformers") transformers_logger.setLevel(logging.MANINO) # Proparing Truin data train_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) test_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) test_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) test_dfi = pd.read_csv('/home/ubuntu/d2_test_csv', error_bad_lines=False) test_dfi = pd.read_csv(-trains_store = 1000 model_args5.newline_defines = True model_args5.metrine_sinpotchs = 3 model_args5.metrine_sinpotchs = True model_args5.metrine_sinpotchs = 3 model_args5.secalust_input.test = read_invertine_sinder_indes add_args5.secalust_input.test = read_invertine_sinder_indes add_args5.secalust_input.test = read_invertine_sinder_indes add_args5.secalust_input.tes = True model_args5.secalust_input.test = secalust model_args5.secalust_input.test = secalust model_args5.secalust_input.test = secalust model_args5.secalust</pre>	1:	import wandb
True import torch igit clone https://github.com/NVIDIA/apex cd apex igit clone https://github.com/NVIDIA/apex cd apex ipip install -vno-cache-dirglobal-option="cuda_ext" ./ import pandsa as pd from sharn.metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel logging.bssicConfig(level=logging.INFO) transformers_logger = logging.getLogger('transformers") transformer_logger.stutevel(logging.IAMON) transformer_logger.stutevel(logging.IAMON) transformer_logger.stutevel(logging.IAMON) transformer_logger.stutevel(logging.IAMON) determing train data train.dfi = pd.read_csv('home/ubuntu/d2_train.csv', error_bad_lines=False) vel.dfi = pd.read_csv('home/ubuntu/d2_train.csv', error_bad_lines=False) # Args for model model_args5.evaluat_during_training = True model_args5.evaluat_during_training = True model_args5.evaluat_during = True model_args5.evaluat_during = True model_args5.evaluat_during = True model_args5.evaluat_during_training = True model_args5.evaluat_during = True model_args5.evaluat_during_training = True model_args5.evaluat_during_training = True model_args5.evaluat_during = True model_args5.evaluat_during = True model_args5.evaluat_during_training = True model_args5.evaluat_du	1:	wandb.login()
<pre>import torch igit clone https://github.com/NVIDIA/apex igit clone https://github.com/NVIDIA/apex id apex iigit clone https://github.com/NVIDIA/apex id apex iigit clone https://github.com/NVIDIA/apex id apex iigit clone https://github.com/NVIDIA/apex iigit clone clon</pre>		True
<pre>igit clone https://github.com/NVIDIA/apex igit clone https://github.com/NVIDIA/apex id dapex igit clone https://github.com/NVIDIA/apex import logging import logging import pandss as pd from sklearn.metrics import accuracy_score from sklearn.metrics import accuracy_score from sklearn.metrics.import accuracy_score from sklearn.metrics.classification import classificationArgs, classificationModel logging.basicConfig(level=logging.igtlogger("transformers") transformers_logger.setLevel(logging.WANING) # Preparing train dat train_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) test_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) test_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args5.evaluate_during_training_stops = 1000 model_args5.evaluate_during_training_stops = 1000 model_args5.sublate_during_training_stops = 1000 model_args5.sublate_diff_i = rrue model_args5.sublate_diff_i = rrue model_args5.sublate_diff_i = rrue model_args5.sublate_diff_i = r/mom/ubuntu/outputs/best_model' model_args5.sublate_diff = r/mom/ubuntu/outputs/output model_args5.subput_diff = rrue model_args5.subput_diff = rrue model_args5.subput_diff = rect model_args5.subput_diff br/>model_args5.subput_diff model_aff, eval_diff= rue rect model_args5.subput_diff model_aff= classificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # from the model model_aff, eval_diff= accuracy_score) # from the model</pre>	1:	import torch
<pre>i cd apex i cd apex i pip install -vno-cache-dirglobal-options"cpp_ext"global-options"cuda_ext" ./ i import logging import pandss as pd from sklearn.mtrics import accuracy_score from singletransformers.classification import classificationArgs, classificationModel logging.basicConfig(level=logging.igtlogger("transformers") transformers_logger = logging.igtlogger("transformers") transformers_logger.setLevel(logging.WANING) # Preparing train dat train_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) val_dfi = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) test_dfi = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args5.evaluate_during_training_stops = 1000 model_args5.evaluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_diri = '/home/ubuntu/duruty/best_model' model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_training_stops = 1000 model_args5.waluate_during_traine_during_traine_during_traine_during_traine_during_traine_during_traine_during_traine_during_traine_during_traine_durin</pre>	1:	!git clone https://github.com/NVIDIA/apex
<pre>[pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./ import logging import pandas as pd from sklearn.metrics import accuracy_score from simpletransformers.classification import classificationArgs, classificationModel logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger('transformers") transformers_logger.setLevel(logging.WARNING) # Preparing train dat train_dfl = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_dfl = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_dfl = pd.read_csv('/home/ubuntu/d2_text.csv', error_bad_lines=False) model_argsS.cmal_batch_size = 54 model_argsS.cmal_batch_size = 54 model_argsS.cmal_batch_size = 17ue model_argsS.cent_dfl = */home/ubuntu/outputs/cache_dir' model_argsS.cent_dfl = */home/ubuntu/outputs/cache_dir' model_argsS.cent_dfl = */mome': "BertMin2' model_argsS.centy_stopping metric = *mec' model_argsS.centy_stopping metric</pre>	1 : [	cd apex
<pre>import logging import pandas as pd from skilearn.metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger("transformers") transformers_logger.setLevel(logging.WARNING) # Preparing train data train_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) test_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args5 = classificationArgs() model_args5.evaluate_during_training = True model_args5.evaluate_during_training = 1000 model_args5.evaluate_during_training = 1000 model_args5.train_batch_size = 64 model_args5.train_batch_size = 64 model_args5.strain_doch_dir = '/home/ubuntu/outputs/output' model_args5.strain_doch_dir = '/home/ubuntu/outputs/output' model_args5.strain_doch_dir = '/home/ubuntu/outputs/output' model_args5.strain_dock_dir = '/home/ubuntu/outputs/output' model_args5.strain_gput = 2 model_args5.strain_gput = 2 model_args5.strain_gput = 2 model_args5.strain_gput = te-5 model_args5.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _ess.strain_model( train_df1 _est_args5.esm/df1</pre>	1:	<pre>!pip install -vno-cache-dirglobal-option="cpp_ext"global-option="cuda_ext" ./</pre>
<pre>logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger("transformers") transformers_logger.setLevel(logging.WARNING) # Preparing train data train_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_eval.csv', error_bad_lines=False) test_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args5 = ClassificationArgs() model_args5.evaluate_during_training = True model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.vandus_sed = 4 model_args5.vandus_sed = 4 model_args5.vandus_dif = 235 model_args5.nam_biste = 64 model_args5.verwrite_output_dfa = True model_args5.verwrite_output_dfa = True model_args5.verwrite_output_dfa = True model_args5.cach_dir = '/home/ubuntu/outputs/best_model' model_args5.verwrite_output_dfa = "ree model_args5.verwrite_output_dfa = "ree model_args5.verwrite_output_dfa = "ree model_args5.verwrite_output_dfa = "BertMin2") model_args5.verwrite_output_dfa = "nee" model_args5.verwrite_opping_metric="mace" model_args5.verwrite_opping = True model_args5.verwrite_opping = True model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" model_args5.early_stopping_metric="mace" # Create a TransformerModel model_args5.early_stopping_metric="mace" # Create a TransformerModel model_args5.early_stopping_metric="mace" # Create a TransformerModel model_args5.early_stopping_metric="mace" # Create a TransformerModel model_args5.early_stopping_metric="mace" # Create a TransformerModel model_args5.early_stopping_metric="mace" # Create a TransformerModel # Train_dfi_ # eval_df=val_df1, # eval_df=val_df1, # acc</pre>	1 =	<pre>import logging import pandas as pd from sklearn.metrics import accuracy_score from simpletransformers.classification import ClassificationArgs, ClassificationModel</pre>
<pre># Preparing train data train_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_eval.csv', error_bad_lines=False) test_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False) # Args for model model_args5 classificationArgs() model_args5.evaluate_during_training = True model_args5.evaluate_during_training = True model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.vevaluate_during_training_steps = 1000 model_args5.vevaluate_during_training_steps = 1000 model_args5.vevaluate_during_traine_steps = 3 model_args5.vevaluate_during_traine_evaluate=during_traine_evaluate=during_traine_evaluate=during_traine_evaluate=during_traine_evaluate=during_traine_evalua</pre>		<pre>logging.basicConfig(level=logging.INFO) transformers_logger = logging.getLogger("transformers") transformers_logger.setLevel(logging.WARNING)</pre>
<pre># Args for model model_args5 = ClassificationArgs() model_args5.exaluate_during_training = True model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.evaluate_during_training_steps = 1000 model_args5.max_seq_length = 235 model_args5.max_seq_length = 235 model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = '/home/ubuntu/outputs/best_model' model_args5.overwrite_output_dir = '/home/ubuntu/outputs/cache_dir' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.s.ngpu = 2 model_args5.s.ngpu = 2 model_args5.s.ngpu = 2 model_args5.s.ngpu = 2 model_args5.s.early_stopping_metric = "mcc" model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric = "mcc" model_args5.do_lower_case = True # Create a TransformerModel model['bert', 'google/bert_uncased_L-12_H-512_A-8'', use_cuda=True, num_labe] # Train the model model[ model_afs = ClassificationModel('bert', ''google/bert_uncased_L-12_H-512_A-8'', use_cuda=True, num_labe] # Create a ClassificationModel(''bert'', ''''''''''''''''''''''''''''''''''</pre>		<pre># Preparing train data train_df1 = pd.read_csv('/home/ubuntu/d2_train.csv', error_bad_lines=False) eval_df1 = pd.read_csv('/home/ubuntu/d2_eval.csv', error_bad_lines=False) test_df1 = pd.read_csv('/home/ubuntu/d2_test.csv', error_bad_lines=False)</pre>
<pre>model_argss = ClassificationArgs() model_argss.evaluate_during_training = True model_argss.evaluate_during_training = True model_argss.evaluate_during_training = True model_argss.eval_batch_size = 32 model_argss.eval_batch_size = 54 model_argss.num_train_epochs = 3 model_argss.overwrite_output_dir = True model_argss.eval_batch_size = 64 model_argss.eval_batch_size = 74 model_argss.condel_dir = '/home/ubuntu/outputs/best_model' model_argss.conde_dir = '/home/ubuntu/outputs/cathe_dir' model_argss.cache_dir = '/home/ubuntu/outputs/cathe_dir' model_argss.cache_dir = '/home/ubuntu/outputs/cathe_dir' model_argss.cache_dir = '/home/ubuntu/outputs/cathe_dir' model_argss.cache_dir = '/home/ubuntu/outputs/cathe_dir' model_argss.output_dir = mame": "BertMin2"} model_argss.icache_dir = '/nome'ris = "met" model_argss.early_stopping metric = "incc" model_argss.early_stopping_metric = "incc" model_argss.early_stopping_metric = false model_argss.learning_rate = 4e-5 model_argss.do_lower_case = True # Create a TransformerModel model(     train_dfi,     eval_df=eval_dfi,     accuracy= accuracy_score) # Environmedel # Train the model # Final with e model # Create a councel #</pre>		# Args for model
<pre>model_args5.evaluate_during_training_steps = 1000 model_args5.eval_batch_size =32 model_args5.eval_batch_size =32 model_args5.num_train_epochs = 3 model_args5.num_train_epochs = 3 model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = True model_args5.overwrite_output_dir = i/home/ubuntu/outputs/best_model' model_args5.cval_dir = '/home/ubuntu/outputs/output' model_args5.cval_dir = '/home/ubuntu/outputs/cache_dir' model_args5.wandb_project = "Msc" model_args5.use_eanly_stopping = True model_args5.evaly_stopping = True model_args5.evaly_stopping_metric = "mcc" model_args5.evaly_stopping_metric = 5 model_args5.dolower_case = True # Create a TransformerModel model("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model( train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5 = ClassificationArgs() model_args5.evaluate_during_training = True</pre>
<pre>model_args5.eval_batch_size =32 model_args5.max_seq_length = 235 model_args5.max_seq_length = 235 model_args5.rnan_batch_size = 64 model_args5.num_train_epochs = 3 model_args5.overwrite_output_dir = True model_args5.everwrite_output_dir = True model_args5.output_dir = '/home/ubuntu/outputs/best_model' model_args5.cach_dir = '/home/ubuntu/outputs/best_model' model_args5.cach_dir = '/home/ubuntu/outputs/cach_dir' model_args5.cach_dir = '/home/ubuntu/outputs/cach_dir' model_args5.son_gpu = 2 model_args5.son_gpu = 2 model_args5.son_gpu = 2 model_args5.son_gpu = 2 model_args5.son_gpu = 2 model_args5.son_gpu = 7rue model_args5.son_gpu = 7rue model_args5.son_gpu = 7rue model_args5.son_long_metric = "mcc" model_args5.son_long_metric = Talse model_args5.do_lower_case = True # Create a TransformerModel model_ds = 5 train_model( train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # 5 urg/unte the model # Train the model # Train_df1, # Create the model # Train_df1, # Create the model # Create the model # Train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # 5 urg/unte the model # 5 urg/unte the urg/unte # 5 urg/unte the urg/unte # 5 urg/unte the urg/unte # 5 urg/unte # 5 urg/unte # 5 urg/un</pre>		<pre>model_args5.evaluate_during_training_steps = 1000 model_args5.manual_seed = 4</pre>
<pre>model_args.imat_nbatch_size = 64 model_args.train_batch_size = 64 model_args.train_batch_size = 64 model_args.overwrite_output_dir = True model_args.overwrite_output_dir = True model_args.soverwrite_output_dir = ''home/ubuntu/outputs/best_model' model_args.cache_dir = ''home/ubuntu/outputs/cache_dir' model_args.cache_dir = ''home/ubuntu/outputs/cache_dir' model_args.wandb_project = "Msc" model_args.wandb_project = "Msc" model_args.soverwrite_output_dir = True model_args.soverwrite_output_outputs/cache_dir' model_args.soverwrite_output_dir = "BertMin2"} model_args.soverwrite_output_dir = True model_args.soverwrite_output_dir = "mcc" model_args.soverwrite_output_metric = "mcc" model_args.dover_case = True # Create a TransformerModel model_dif = 4e-5 model_args.to_lower_case = True # Train the model model_dif = classificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model # Sover_model_args.sover_dir # Create accuracy_score</pre>		<pre>model_args5.eval_batch_size =32 model_args5.eval_batch_size =32</pre>
<pre>model_args5.num_train_epochs = 3 model_args5.nemtrain_epochs = 3 model_args5.reprocess_input_dit = True model_args5.reprocess_input_data = True model_args5.sest_model_dir = '/home/ubuntu/outputs/best_model' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.wandb_project = "Msc" model_args5.wandb_project = "Msc" model_args5.searly_stopping = True model_args5.early_stopping metric = "mcc" model_args5.early_stopping_metric = Talse model_args5.do_lower_case = True # Create a TransformerModel model_afs = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_aft,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5.max_seq_iength = 255 model_args5.train_batch_size = 64</pre>
<pre>model_args5.reprocess_input_data = True model_args5.reprocess_input_data = True model_args5.sest_model_dir = '/home/ubuntu/outputs/best_model' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.use_early_stopping = True model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric = false model_args5.early_stopping_patience = 5 model_args5.do_lower_case = True # Create a TransformerModel model_afs = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_afs.train_model( train_df1, eval_df=eval_df1, accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5.num_train_epochs = 3 model_args5.num_train_epochs = 3</pre>
<pre>model_args5.best_model_dir = '/home/ubuntu/outputs/best_model' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.cache_dir = '/home/ubuntu/outputs/cache_dir' model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.use_early_stopping = True model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric = False model_args5.do_lower_case = True # Create a TransformerModel model_afs = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_af.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model # Train the model</pre>		model_argsJ.reprocess_input_data = True
<pre>model_args5.cathe_dir = '/home/ubuntu/outputs/cathe_dir' model_args5.cathe_dir = '/home/ubuntu/outputs/cathe_dir' model_args5.wandb_project = "Msc" model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.mgpu = 2 model_args5.use_early_stopping = True model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric = False model_args5.early_stopping_patience = 5 model_args5.learning_rate = 4e-5 model_args5.dollower_case = True # Create a TransformerModel model_afs = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model </pre>		<pre>model_args5.best_model_dir = '/home/ubuntu/outputs/best_model' productions/final content discussion in the production of the producti</pre>
<pre>model_args5.wandb_project = "Msc" model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.use_early_stopping = True model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric_minimize = False model_args5.early_stopping_metric_minimize = False model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True # Create a TransformerModel model_afs = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_afs.train_model( train_df1, eval_df=eval_df1, accuracy= accuracy_score) # Evaluate the model</pre>		model_args5.cotput_air = //home/ubuntu/outputs/outputs
<pre>model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.usplue 2 model_args5.usplue z model_args5.usplug_metric = "mcc" model_args5.early_stopping_metric_minimize = False model_args5.early_stopping_patience = 5 model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True # Create a TransformerModel model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5.wandb_project = "Msc"</pre>
<pre>model_args5.use_early_stopping = True model_args5.use_early_stopping_metric = "mcc" model_args5.early_stopping_metric_minimize = False model_args5.early_stopping_patience = 5 model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True # Create a TransformerModel model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5.wandb_kwargs = {"name": "BertMin2"} model_args5.m epu = 2</pre>
<pre>model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric_minimize = False model_args5.early_stopping_patience = 5 model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True # Create a TransformerModel model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		model_args5.use_early_stopping = True
<pre>model_args5.early_stopping_matterce = 5 model_args5.early_stopping_patience = 5 model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True # Create a TransformerModel model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5.early_stopping_metric = "mcc" model_args5.early_stopping_metric = Telee</pre>
<pre>model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True # Create a TransformerModel model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		model_argsS.early_stopping_motifice = 5
<pre># Create a TransformerModel # Create a TransformerModel model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_args5.learning_rate = 4e-5 model_args5.do_lower_case = True</pre>
<pre># Credies = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_label # Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		# Create a TransformerModel
<pre># Train the model model_a5.train_model(     train_df1,     eval_df=eval_df1,     accuracy= accuracy_score) # Evaluate the model</pre>		<pre>model_a5 = ClassificationModel("bert", "google/bert_uncased_L-12_H-512_A-8", use_cuda=True, num_labels=5</pre>
<pre>train_df1, eval_df=eval_df1, accuracy= accuracy_score) # Evaluate the model</pre>		model_a5.train_model(
accuracy_score)		train_df1, eval df=eval df1.
# Evolution the medal		accuracy= accuracy_score)
<pre>result, model_outputs, wrong_predictions = model_a5.eval_model(test_df1,accuracy=accuracy_score)</pre>		<pre># Evaluate the model result, model_outputs, wrong_predictions = model_a5.eval_model(test_df1,accuracy=accuracy_score)</pre>
<pre>wandb.log({'BertMinD2': result})</pre>		
4		wandb.log({ BertMinD2 : result})

Figure 218: Smaller Bert D2 Code

💭 JupyterLab 🛛 🗙		- σ					
(←) → 健 @		🛛 🔒 https://app.wandb.ai/browner5	i5/Msc/runs/45oa7ezu/overview?workspace=user-browner55			120%	♡ ☆ IN ① @ 🕈 🛃 🔤
111 (	Q browner55 >	Projects $>$ Msc $\triangle$ $>$ Runs $>$ S	maller BERT D2 > Overview				
① Smaller BERT D2 ℓ						:	
Mhat makes this run special? ℓ							
	Privacy						
₹ <sub>mm</sub> ⊧	Tags	+					
s73	Author	A browner55					
>	State	crashed					
~_	Start time	August 12th, 2020 at 6:18:37 pm					
D	Duration	5h 22m 465					
	Run path	ath browner55/Msc/45oa7ezu					
	OS	Uname Gata-msc					
	Python version	version 3.7.6					
	Python executa	Python executable /usr/local/share/anaconda3/bin/python					
	Git repository	git clone https://github.com/	NVIDIA/apex				
Git state     git checkout -b "Smaller-BERT-D2" 5d9b5cbc2fb0e6c525c067123d66f1827440c7a8       Command     cpython with no main file> -f /home/ubuntu/.local/share/jupyter/runtime/kernel-3eac4f75-5096-43bc-9585-b805f1eaa038.json       System Hardware     GPU type       GeForce GTX1080 Ti							
	W&B CLI Version 0.9.4						
	Config paramete	ers describe your model's inputs. <b>L</b>	earn more Page 1 of 4 < >		Summary metrics describe your results. Learn more Q. Search		Pagelofl < >
	Name		Value		Name	Value	
	@ adam_ep	silon	1.000e-8	^	al Training loss	0.3659	0
	ø best_mod	lel_dir	/home/ubuntu/outputs/best_model		and accuracy	0.6788	
	<ul> <li>cache_dir</li> <li>custom_layer_parameters</li> </ul>		/home/ubuntu/outputs/cache_dir		all eval_loss	0.8303	
			0		all global_step	24000	
	custom_p	arameter_groups	0		ad graph_0	graph-file	0
③ dataload		er_num_workers	6	- 1	al Ir	0.000006245	
	Ø do_lower.	_case	true	and moc	0.5986		
	@ early_stopping_consider_epochs		false		all train_loss	0.3659	
	early_stopping_delta		0				0
	early_stopping_metric		mcc				
	early_stop	oping_metric_minimize	false				
	early_stop	oping_patience	5				
	encoding		-				
	eval_batc	h_size	32				
	evaluate	during_training	true				

Figure 219: Smaller Bert D2 Results

## 7 Conclusion

By utilizing the information above the process of implementing the candidates project can be achieved. It should be noted that due to the randomness of Neural Networks the exact results may not be achieved. To control this the candidate used manual seed for the Simple Transformers library however the candidate didn't realize that another library used randomness also. Therefore, there will be a slight difference in the results achieved to the Technical Report when the code is re-run. Finally, the project on Weights and Biases has been made public until the end of September and can be viewed here<sup>15</sup>.

<sup>&</sup>lt;sup>15</sup> <u>https://app.wandb.ai/browner55/Msc?workspace=user-browner55</u>