

N° d'ordre :  
N° de série :

**PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA**  
**Ministry of Higher Education and Scientific Research**



**ECHAHID HAMMA LAKHDAR UNIVERSITY - EL OUED**  
**FACULTY OF EXACT SCIENCES**  
**Computer Science department**



**End of Study Memory**  
**Presented for the Diploma of**

## **ACADEMIC MASTER**

Domain : **Mathematics and Computer Science**  
spinneret: **Computer Science**  
Speciality : **Artificial Intelligence and Distributed Systems**

Presented by :

- **Selmi Mohammed**
- **Soltani Oussama**

### **Theme**

# **Fake News Generation and Detection**

Supported in: 20 - 06 - 2019 In front of jury:

M.	KHOLLADI Nadjoua Houda	MCA	President
M.	KHEBBACHE Mohib Eddine	MAA	Reporter
Dr.	El Moataz Billah Nagoudi	MAA	Supervisor

**University Year: 2019/2020**

# Acknowledgements

First of all, we thank **Allah** for giving us the opportunity to be at this position and providing us with knowledge and patience to finish this work. Second, we would like to express our sincere gratitude to Dr. *Nagoudi El Moatez Billah*, our supervisor for his continued support for research, inquiry, patience, motivation and knowledge. As well as his help and guidance and valuable advice at all time of research and throughout these months of writing this thesis.

In addition to our supervisor, we would like to thank our parents, the friends we had their help. We thank the rest of the discussion committee for their insightful comments, encouragement and advice, and anyone who has had a hand from near or far in helping us even by encouraging and motivating.

# Abstract

Computer science has developed to a great extent, especially in artificial intelligence. Artificial intelligence is a sub-field that has big remarkable progress over the latest years. Artificial intelligence is used to solve many problems by using natural language processing. Natural language processing covers many studies and it is considered The main reason for the advancement of techniques for understanding human behaviors. Natural language processing can be used to solve problems such as plagiarism detection, word, and information extraction from texts, and it is also used in machine translation and text classification. In the Arabic language, there is a lack in Arabic dataset in every domain of artificial intelligence, especially in text classification. Text classification has many types of research and one of those researches is Fake news Detection. This is our problem, in this thesis we will try to solve the problem of lack in Arabic dataset and we will propose two programs that have a difference in programming language and the functions. And we will show the results of it to know how strong are.

**Keywords:** Fake News , Disinformation, Social Media, , Generation.

# Resumé

L'informatique s'est largement développée, en particulier dans l'intelligence artificielle. L'intelligence artificielle est un sous-domaine qui a fait de grands progrès remarquables au cours des dernières années. L'intelligence artificielle est utilisée pour résoudre de nombreux problèmes en utilisant le traitement du langage naturel. Le traitement du langage naturel couvre de nombreuses études et il est considéré comme la principale raison de l'avancement des techniques pour comprendre les comportements humains. Le traitement du langage naturel peut être utilisé pour résoudre des problèmes tels que la détection de plagiat, l'extraction de mots et d'informations à partir de textes, et il est également utilisé dans la traduction automatique et la classification de texte. Dans la langue arabe, il y a un manque dans l'ensemble de données arabe dans tous les domaines de l'intelligence artificielle, en particulier dans la classification de texte. La classification de texte comporte de nombreux types de recherche et l'une de ces recherches est la détection des fausses nouvelles. C'est notre problème, dans cette thèse, nous allons essayer de résoudre le problème du manque de données arabes et nous proposerons deux programmes qui ont une différence dans le langage de programmation et les fonctions. Et nous en montrerons les résultats pour savoir à quel point ils sont forts.

**Mots-clés:** Fake News, Désinformation, Médias Sociaux, Génération.



## ملخص

لقد تطور علم الحواسيب إلى أبعد الحدود ولا سيما في الذكاء الصناعي . إن الذكاء الصناعي هو مجال فرعي قد عرف تقدما ملحوظا خلال السنوات الماضية . يستعمل الذكاء الصناعي في حل العديد من المشاكل باستعمال معالجة اللغة الطبيعية . إن معالجة اللغة الطبيعية تغطي ال كثير من الدراسات وكما تعتبر حجر الأساس في تقدم تقنيات فهم التصرفات البشرية . يمكن استعمال معالجة اللغة الطبيعية في حل مشاكل مثل كشف الانتحال واستخراج الكلمات و المعلومات من النصوص وكما تستخدم أيضا في الترجمة الآلية و تصنيف النصوص .

في اللغة العربية ، هناك نقص في مجموعة البيانات العربية في كل مجال من مجالات الذكاء الاصطناعي وخاصة في تصنيف النص . يحتوي تصنيف النص على الكثير من الأبحاث ومن هذه الأبحاث كشف الأخبار الكاذبة . هذه هي مشكلتنا ، في هذه الرسالة سنحاول حل مشكلة نقص مجموعة البيانات العربية وسنقترح برنامجين يختلفان في لغة البرمجة والوظائف . وسوف نعرض نتائجها لنعرف مدى قوتها .

**الكلمات المفتاحية :** أخبار مزيفة ، معلومات مضللة ، مواقع التواصل الاجتماعي ، توليد .

# Contents

<b>Introduction</b>	<b>1</b>
<b>1 State of the Art</b>	<b>2</b>
1.1 Artificial Intelligence . . . . .	3
1.1.1 Definition . . . . .	3
1.2 Machine Learning . . . . .	3
1.2.1 Definition . . . . .	3
1.3 Word Embedding . . . . .	3
1.3.1 Definition . . . . .	3
1.3.2 Field Of Use . . . . .	4
1.3.3 Different Architectures . . . . .	5
1.3.4 Continuous Bag Of Word Model (CBOW) . . . . .	5
1.4 Language Model . . . . .	7
1.4.1 Definition . . . . .	7
1.4.2 Bert . . . . .	7
1.4.3 GPT-2 (Text Generation) . . . . .	8
1.4.4 Grover Model . . . . .	15
1.5 Related Work . . . . .	16
1.5.1 Lair Liar Pants On Fire . . . . .	17
1.5.2 Detecting Deceptive Reviews using Generative Adversarial Networks	20
1.5.3 KnowledgE-Based Fake News Detection . . . . .	21
<b>2 Collecting And Generating Datasat</b>	<b>22</b>
2.1 The Steps Of Collecting Data . . . . .	23
2.1.1 Collecting List of Arabic Newspapers . . . . .	23
2.1.2 Filtering The Newspapers . . . . .	23
2.1.3 The Write Of Code . . . . .	23
2.1.4 The Process Of Crawl . . . . .	24
2.2 The Problem Faced Us . . . . .	24
2.3 Data Analysis . . . . .	25
2.4 Data Statistics . . . . .	26
2.5 Generate Dataset . . . . .	29
2.5.1 Arabic Pre-Trained GPT-2 . . . . .	30
2.6 The Problems of Arabic Pre-Trained GPT-2 . . . . .	33
2.7 Solve The Arabic Pre-Trained GPT-2 Problems . . . . .	33
2.7.1 Arabic Pre-Trained GPT-2 ++ . . . . .	33
2.8 Test Arabic Pre-Trained GPT-2 and Arabic Pre-Trained GPT-2 ++ . . .	34
2.8.1 Questionnaire . . . . .	34

2.8.2	The Result . . . . .	35
2.8.3	Examples . . . . .	35
<b>3</b>	<b>Fake News Detector System</b>	<b>37</b>
3.1	CBOW Based On Java System . . . . .	38
3.1.1	Training information of CBOW Based On Java System . . . . .	38
3.1.2	Test Phase . . . . .	38
3.1.3	The Class Of Java Project . . . . .	39
3.1.4	CBOW Based On Java System Results . . . . .	45
3.2	Detector System Based On Bert . . . . .	47
3.2.1	Training information of Detector System Based On Bert . . . . .	47
3.2.2	The implementation of Bert system for classify the news . . . . .	48
3.2.3	Detector System Based On Bert Results . . . . .	49
	<b>Conclusion</b>	<b>57</b>
	<b>bibliography</b>	<b>58</b>



# List of Figures

1.1	Relations between words according to word embeddings <a href="#">Levy and Goldberg (2014a)</a> . . . . .	4
1.2	Using STS on chat conversation <a href="#">Yinfei Yang (2018)</a> . . . . .	4
1.3	.5figure.1.3	
1.4	One-Word Context <a href="#">Alammar (2019)</a> . . . . .	6
1.5	Multi-Word Context <a href="#">Alammar (2019)</a> . . . . .	6
1.6	versions of GPT-2 <a href="#">Alammar (2019)</a> . . . . .	9
1.7	GPT-2 predict next word <a href="#">Alammar (2019)</a> .. . . .	9
1.8	Positional Encoding + Input Embedding <a href="#">Alammar (2019)</a> . . . . .	10
1.9	the path of token in GPT-2 <a href="#">Alammar (2019)</a> . . . . .	10
1.10	Dot product between vector output and matrix embedding <a href="#">Alammar (2019)</a> . . . . .	11
1.11	Matrix of probability of tokens. . . . .	11
1.12	creating Q, V, K vectors from 3 matrix <a href="#">Alammar (2019)</a> . . . . .	11
1.13	calculassions of score <a href="#">Alammar (2019)</a> . . . . .	12
1.14	sum all vector to obtain Z <a href="#">Alammar (2019)</a> . . . . .	12
1.15	masked self and self attention <a href="#">Alammar (2019)</a> . . . . .	13
1.16	Multi-Head Attention -12 layers <a href="#">Alammar (2019)</a> . . . . .	13
1.17	Merging all Heads. . . . .	14
1.18	result of dot product between Z and all head. . . . .	14
1.19	layer 1 <a href="#">Alammar (2019)</a> . . . . .	15
1.20	layer 2 <a href="#">Alammar (2019)</a> . . . . .	15
1.21	Architecture of LSTM <a href="#">Phi (2018)</a> . . . . .	18
1.22	Forget gate and it output <a href="#">Phi (2018)</a> . . . . .	18
1.23	update gate architecture and it outputs <a href="#">Phi (2018)</a> . . . . .	19
2.1	hugging face web page of gpt-2 small Arabic <a href="#">AKHOLI (2020)</a> . . . . .	30
2.2	Install Arabic Pre-Trained GPT-2 . . . . .	31
2.3	Code Arabic Pre-Trained GPT-2 . . . . .	31
2.4	An example of generating sentences using Arabic Pre-Trained GPT-2 . . . . .	32
2.5	Code this algorithm for cleaning . . . . .	34
2.6	Example superiority of Arabic Pre-Trained GPT-2 ++ of Arabic Pre-Trained GPT-2 . . . . .	36
3.1	Method Initweights . . . . .	39
3.2	ActFunc . . . . .	40
3.3	CalcGrads . . . . .	41
3.4	ElementWiseDrive . . . . .	41
3.5	InitStochGradDiscen . . . . .	42
3.6	initAccumulVals . . . . .	42

3.7	LossFunction . . . . .	42
3.8	DLossFunc . . . . .	42
3.9	UpdatValsOf . . . . .	43
3.10	UpdatGradsOf . . . . .	43
3.11	trinWithDSL . . . . .	43
3.12	train . . . . .	44
3.13	Fake Sentence, output CBOW Based On Java System Fake, output Person Fake . . . . .	45
3.14	Fake Sentence, output CBOW Based On Java System Real, output Person Fake . . . . .	46
3.15	Fake Sentence, output CBOW Based On Java System Real, output Person Real . . . . .	46
3.16	Fake Sentence, output CBOW Based On Java System Fake, output Person Real . . . . .	47
3.17	Install and Import Ktrain . . . . .	48
3.18	Splitting DataSet . . . . .	48
3.19	Loading And Using Pre-Train Model . . . . .	49
3.20	Prepare To Predict And Save . . . . .	49
3.21	Load and Use our system to predict . . . . .	49
3.22	Fake Sentence, output Bert 210,000 Sentences 10 Epochs Fake, output Person Fake . . . . .	50
3.23	Fake Sentence, output Bert 210,000 Sentences 10 Epochs Real, output Person Fake . . . . .	51
3.24	Fake Sentence, output Bert 210,000 Sentences 10 Epochs Real, output Person Real . . . . .	51
3.25	Fake Sentence, output Bert 210,000 Sentences 10 Epochs Fake, output Person Real . . . . .	52
3.26	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Person Fake . . . . .	52
3.27	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Person Fake . . . . .	53
3.28	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Person Real . . . . .	53
3.29	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Person Real . . . . .	54
3.30	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Bert 210,000 Sentences 10 Epochs Fake . . . . .	54
3.31	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Bert 210,000 Sentences 10 Epochs Fake . . . . .	55
3.32	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Bert 210,000 Sentences 10 Epochs Real . . . . .	55
3.33	Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Bert 210,000 Sentences 10 Epochs Real . . . . .	56

# List of Tables

1.1	The evaluation results on the LIAR dataset.The top section: text-only models.The bottom: text + meta-data hybrid models <a href="#">Wang (2017)</a> . . . .	20
2.1	Country and domains. . . . .	25
2.2	Newspaper and Country. . . . .	26
2.3	Average number of words. . . . .	27
2.4	Average number of characters. . . . .	28
2.5	The size of the data. . . . .	29
2.6	The Result . . . . .	35
3.1	Training information of CBOW Based On Java System . . . . .	38
3.2	CBOW Based On Java System Results . . . . .	45
3.3	Training information of Detector System Based On Bert . . . . .	47
3.4	Detector System Based On Bert Results . . . . .	50

# Introduction

Fake news has become a part of the world, and it has become necessary to label every story either fake or legitimate. It is essential that we dig into fake news to protect people from believing false news.

In 21 century not only do mankind's use disinformation to achieve and spread propaganda but even more. Machines and computers are using artificial intelligence to generate dozens of fake stories in the whole world after a few seconds. The language model is the main tool that allows computers and machines to be able to virtualized the language of humans and create a lot of fake data. It is very difficult to Verification the source of content. Artificial Intelligence provides certain features which help us rate the news for authenticity and define it as fake. The best way to combat fake news is using an automated tool.

Artificial intelligence provides us with natural language processing. Fake news detection is one of the important major problems that must be found solutions and algorithms for it. This domain is new, especially in Arabic language. The problem in this domain is suffering from the lack of Arabic dataset and the difficulty dealing with the Arabic language itself and that due to the morphological of Arabic language. There is another problem with a lack of research in this domain. So to solve those problems we wrote this thesis and we are going to talk about:

In this topic we have three chapters: chapter 1 speaks about some of the previous state of the arts that used to generate disinformation and fake news in any human's language. In chapter 2 we are going explain the steps of collecting Arabic Dataset and then explain the tools for that. Also in the same chapter we will mention some problems with Arabic crawling and talk about data analysis, and using one of the tools of language modeling to generate fake dataset. In Final Chapter we are going to define one of the techniques that is used to fight fake news and classify it using deep learning algorithms and also we are going to talk about java system and the implementation of the Bert system for classifying our collected Arabic dataset and compare the results.

# Chapter 1

## State of the Art

*"What's special about human language? It's the most important distinctive human characteristic, the only hope for explainable intelligence and it is a social system."*

*Christopher Manning*

# Introduction

Natural Language Processing and also known as (NLP) [Manning and Schutze \(1999\)](#) is considered as one of the oldest domains in the field of Artificial Intelligence (AI) [Russell and Norvig \(2002\)](#), thus, its history has grown up and gotten wider and deeper. In NLP the basic units are words and the implementation of any system that processes a language depends and relies on words or tokens in its work [Bird et al. \(2009\)](#).

Empirical in this thesis we will take one of a major problems in NLP and explain its techniques and the sub-field and tools of it, the problem is text generation. First of all we define Language Model (LM) in small lines and present brief about it, then we will talk about Text Generation [1.4.3](#) and the tools of the Models (Bert [1.4.2](#), GPT-2 [1.4.3](#) and Grover [1.4.4](#)) that can Generates Texts.

## 1.1 Artificial Intelligence

### 1.1.1 Definition

Artificial intelligence (AI) as a discipline can be considered to be a type of ‘philosophical engineering’. That means AI is a process of taking philosophical ideas and transforming it to an algorithm and implementing them to achieve a certain purpose [Russell and Norvig \(2002\)](#).

That purpose could be anything such as : controlling of agents or use it in semantic web, robotic implementation too [Nielsen \(2015\)](#).

## 1.2 Machine Learning

### 1.2.1 Definition

Machine learning (LM) is a subfield of Artificial intelligence. It allows us to extract knowledge from the data and help to understand the contexts. Mankind achieves many developments in Artificial intelligence (AI) using machine learning [Müller et al. \(2016\)](#). It is a research field that depends and relies on the intersection and utilization of statistics, AI, and computer science and also known as predictive analytics or statistical learning [Michie et al. \(1994\)](#).

## 1.3 Word Embedding

### 1.3.1 Definition

In Natural Language Processing the word is the important unit in the sentence. The human can understand the word and the language but the computer did not. So to make computer understand the language of human it must convert the word into a numerical representation [Levy and Goldberg \(2014a\)](#).

Word Embedding is a representation of the word in the reference .every word in the reference is a vector. The location and the distance between the vectors of word represent the semantic and the likeness of meaning between the words [Levy and Goldberg \(2014a\)](#).

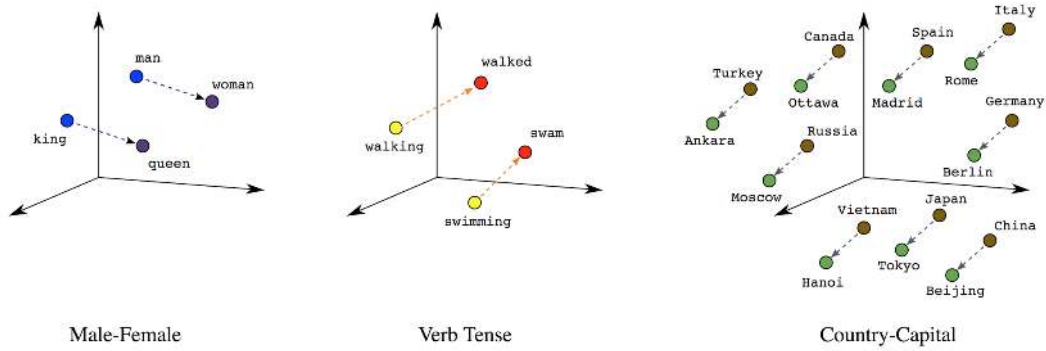


Figure 1.1: Relations between words according to word embeddings [Levy and Goldberg \(2014a\)](#).

### 1.3.2 Field Of Use

In Natural Language Processing the word is not just considered as units in the sentences but it took the syntax and the relationship between each word. Below here is some of utilizations of Word Embedding [Levy and Goldberg \(2014b\)](#):

#### 1.3.2.1 Semantic Textual Similarity

This method is used on a set of documents or terms, the idea of it is use the distance between items just to express on similarity of meaning and semantic contents. In other word the semantic text see how two a couple of items are similarity in linguistic [Levy et al. \(2015\)](#).



Figure 1.2: Using STS on chat conversation [Yinfei Yang \(2018\)](#).

#### 1.3.2.2 Plagiarism Detection

After appearance of internet and the information has available to all people in the world, the bad use and Scientific theft is spread [Khorsi et al. \(2018\)](#), in research field any one can re-use the hard work of others willingly without citing or mentioning them. In all that the crewmate of Jeremy Ferrero has develop a plagiarism detector based on Zord Embedding [Ferrero et al. \(2017\)](#):

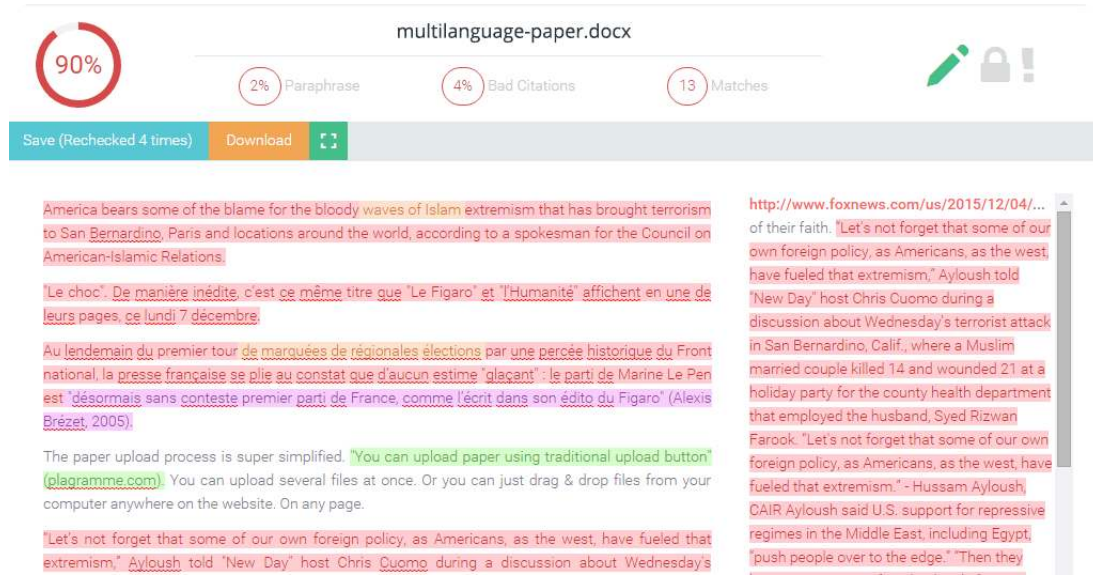


Figure 1.3: Screenshot for an online plagiarism detection tool <sup>1</sup>.

### 1.3.3 Different Architectures

#### 1.3.3.1 Mnih and Hinton (2009)

This work criticized the use of Neural Probabilistic Language Models (NPLMs) in the latest ten years ago and the reason of that is the training process take a lot of the time and expensive cost and that it resulted a less accurate model compared to its non-hierarchical ancestor [Levy and Goldberg \(2014b\)](#).

### 1.3.4 Continuous Bag Of Word Model (CBOW)

CBOW contains two versions, the first is a straightforward version presented in Mikolov et al., Which is only one word in context, while the second version, a multi-word context [Kenter et al. \(2016\)](#).

#### 1.3.4.1 One-Word Context

The simplest version of the Continuous Word Bag (CBOW) model presented in Mikolov et al. (2013a). We assume that only one word is taken into consideration in the context, meaning that the model will predict a target word from the contextual word. The following Figure illustrates the grid model under the simplified context definition [Kenter et al. \(2016\)](#). In context, the vocabulary size is  $V$  and the size of the hidden layer is  $N$  [Wang et al. \(2017\)](#).



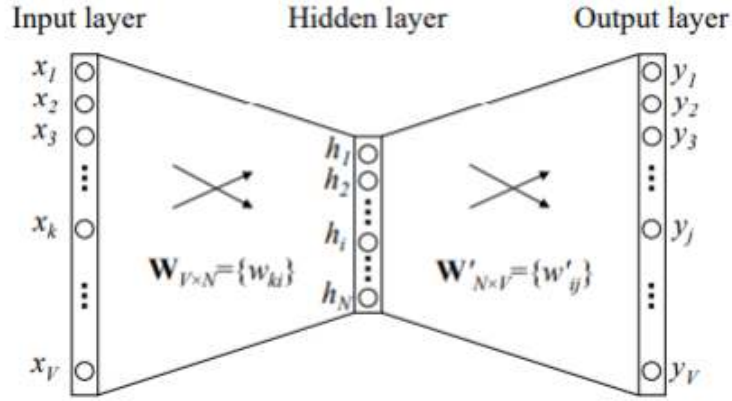


Figure 1.4: One-Word Context [Alammar \(2019\)](#).

The units on adjacent layers are fully connected. The input is a one-hot Encoded Vector, which means for a given input context word, only one out of  $V$  units,  $x_1, x_V$ , will be 1, and all other units are 0 [Wang et al. \(2017\)](#).

#### 1.3.4.2 Multi-Word Context

The figure shows the CBOW model with a multi-word context parameter. When calculating the masked layer output, instead of directly copying the input vector of the input context word [Kenter et al. \(2016\)](#), the CBOW model takes the average of the input context word vectors and uses the product of the weighting matrix hidden weight and mean vector, as output where  $C$  is the number of context words,  $w_1, \dots, w_C$  are the context words and  $vw$  the input vector of a word  $w$  [Wang et al. \(2017\)](#).

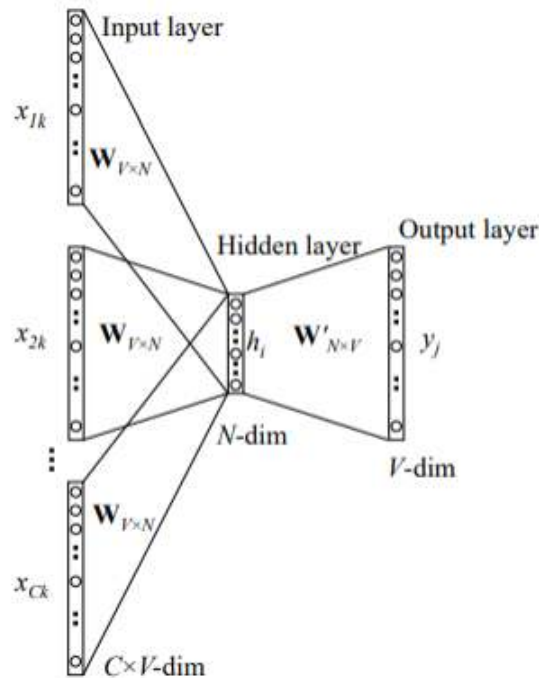


Figure 1.5: Multi-Word Context [Alammar \(2019\)](#).

## 1.4 Language Model

### 1.4.1 Definition

Language Model (LM) contains various algorithms that generate language. Actually the LM created to virtualize human speech and the result of using it is incredible, it is still hard to understand due to the complex architecture of human's brain. There are many LM such as: summarization of texts and Text Generation and it has many types [Bengio et al. \(2003\)](#).

### 1.4.2 Bert

#### 1.4.2.1 Definition

BERT (Bidirectional Encoder Representations from Transformers) is a recent Model Language published by Google AI Language researchers, trained on a large number of corpus for yield good results and achieving accuracy [Sanh et al. \(2019\)](#).

The model has developed many tasks for NLP by developing results in: answering questions, inference of Natural Language, and other domains in NLP. The bidirectional features in this model helps for processing data textual in a short time [Lee and Hsiang \(2019\)](#).

Bidirectional processes the sentence in two directions (right, left) and that allows Bert to understand the meaning of every token [Devlin et al. \(2018\)](#).

#### 1.4.2.2 Other Utilize Of Bert

People used many models to achieve their goals but there are many problems that always face them, such as lack in datasets and effective of the model itself, so to solve these problems they use BERT in:

- **Sentiment analyze** Sentiment analysis is one of major problems in deep learning, the goal that they want to achieve from this problem is to classify texts and know the emotion of humans when they write the text (angry, happy, sad, or normal state) [Alessia et al. \(2015\)](#).

Sometimes sentiment analyzes express opinions from a set of humans. And here are some projects that use sentiment analysis. The name of the project in this work is "HULMUNA", they used Bert model and 4 datasets on Arabic text to classify them. For more details [Press here](#).

- **Machine translation And Image To Iext** In Deep Learning Machine Translation is a process of translating a segment of text from a certain language to another language (target language). It became a main problem that made the researchers think how to optimize it and how to obtain a good translation, that due to increasing connections between humans in the entire world. Bert model was one of the tools used to achieve the work and here are some projects that use In this project they used Bert pre-trained model for neural translation Bert in translation [Press here](#).

### 1.4.2.3 Bert Components

There are many Bert versions but every version has the same architecture and components. In Bert there are three main components: Transformer, Encoder and Decoder [Vaswani et al. \(2017\)](#).

In the Bert model, the transformer block contains Encoder and Decoder but actually Bert used only an Encoder [Devlin et al. \(2018\)](#).

- **Encoder** The encoder is composed of a stack of  $N=6$  identical layers. Each layer has two sub-layers. The first is a multi-head Self-Attention and the second is a fully connected feed-forward network followed by layer normalization [Vaswani et al. \(2017\)](#).
- **Decoder** Decoder has the same identical  $N=6$  layers. And also has the same sub-layers of the encoder the difference is that the decoder has an extra sub-layer in it called Masked-Head Attention and the output of each sub-layer is the input of the function layer-normalization [Vaswani et al. \(2017\)](#).

## 1.4.3 GPT-2 (Text Generation)

### 1.4.3.1 Definition

In NLP Text Generation is an implementation of making sentences and texts by generating sequences (word by word), it uses the statistic for that and it extracts words in probability distribution by knowing previous tokens. In Text Generation there are many tools to generate text and it requires [Shi et al. \(2018\)](#). GPT-2 is one of the Language Models that contains complexity algorithms proposed by OPENAI. This model is different in its function from other language models, it can generate a large number of texts started by one or two tokens [Radford et al. \(2019\)](#).

### 1.4.3.2 Description of GPT-2

GPT-2 Model was created to be a sophisticated technique that can generate a large corpus beginning with one or two tokens, it has many characters and various type architectures [Radford et al. \(2019\)](#).

- **Version of GPT-2** GPT-2 has many version starts from small to bigger:
  - **GPT-2 Small:** has 117M parameters and 12 Layers.
  - **GPT-2 Medium:** has 345M parameters and 24 Layers.
  - **GPT-2 Large:** has 1542M parameters and 46 Layers.
  - **GPT-2 Super Large:** has 1542M parameters and 48 Layers.



Figure 1.6: versions of GPT-2 [Alammar \(2019\)](#).

- Training Data Set

Many of Language Model trained on small corpus in datasets that's why the results and its effectiveness will be less as what they expected, GPT-2 is different in its training phase, it take a various numbers of large corpus in different type of texts and even books to optimize the results of its training [Radford et al. \(2019\)](#).

The datasets used to train GPT-2 is the same dataset used to train Grover models in addition to WebText too [Radford et al. \(2019\)](#).

### 1.4.3.3 GPT-2 Architecture And General Functions

In Architecture of GPT-2 there are only decoder layers. The model has one input token for each word in the sentence, and every token will be stored in its location and Then it will be treated in all the layers, the result of processing the vector through all layers is stored in a file called vocabulary. Then if it come to prediction phase it select the highest probability [Radford et al. \(2019\)](#).



Figure 1.7: GPT-2 predict next word [Alammar \(2019\)](#)..

- Dipper Look Inside the Model [Radford et al. \(2019\)](#)

- **Positional Encoding :** The addition of positional Encoding is very important in the model. It allows the model how the number of tokens in the sentence

that it entered to be processed and what the order of the recent token which must be converted into a vector.

In this phase the input or a sentence will transform into a vector by embedding and mix with the Positional Encoding.



Figure 1.8: Positional Encoding + Input Embedding [Alammar \(2019\)](#).

After mixing the Vector Embedding input tokens with Positional Encoding the next step is when the mixed result goes to the first block in the model to start the operation of the process [Radford et al. \(2019\)](#).

The first block in the model contains on processing the words by sending it to the self-attention process then the second step is sending the vector to its neural network layer. In the transformer architecture the first block processes the word and it sends the resulting vector to the next block to process it. All of the block contain identical processing [Radford et al. \(2019\)](#).



Figure 1.9: the path of token in GPT-2 [Alammar \(2019\)](#).

- **Self –Attention:** There is a problem with the relation between the words in a sentence so for example in this sentence: “ the human can see the objects with his eyes ” The machine can’t find or know the relationship with the word human and his . That is why we must use self-attention. We will explain how it works later [Radford et al. \(2019\)](#).

So after the input goes on self attention and neural network block .The output of it is a multiplication between it and the token embedding. The next image expresses what happened.

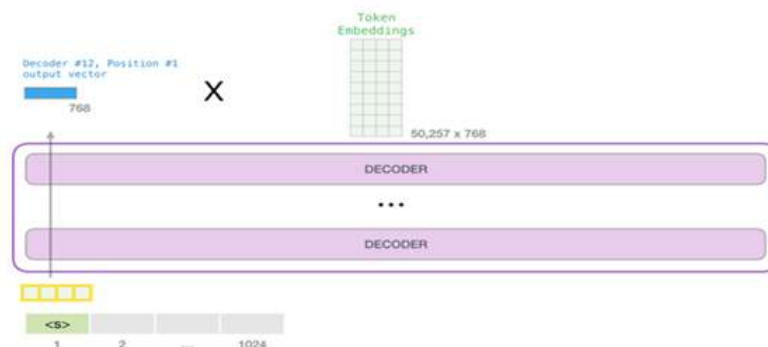


Figure 1.10: Dot product between vector output and matrix embedding [Alammar \(2019\)](#).

The result of a dot product is a matrix of probabilities. 40 top-k token's probabilities are selected to use in prediction.

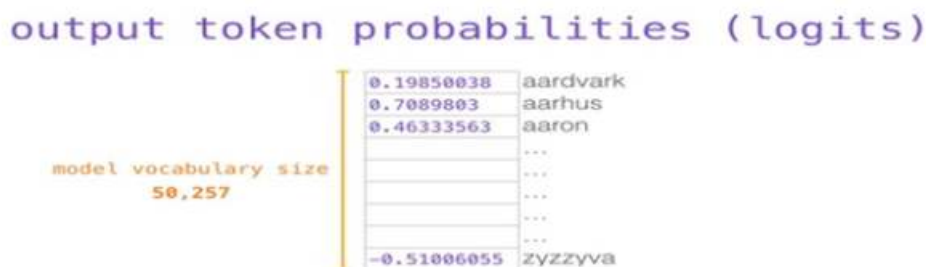


Figure 1.11: Matrix of probability of tokens.

#### – How the self-attention work?

As we said above, the self attention matches between the relations of words (token) in the sentence [Radford et al. \(2019\)](#).

- \* Self attention has 3 weighted matrix  $W_v$ ,  $W_q$ ,  $W_k$
- \* First we must create the vectors :  $q$ ,  $v$ ,  $k$  = query, value, key.
- \* By multiplying the weight matrix with the input token

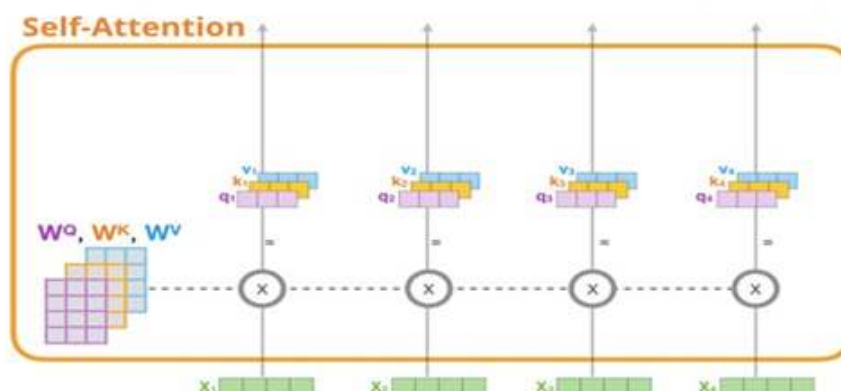


Figure 1.12: creating  $Q$ ,  $V$ ,  $K$  vectors from 3 matrix [Alammar \(2019\)](#).

#### 1. Calculating score

The score calculated by multiplying query vector with all keys of all input vector

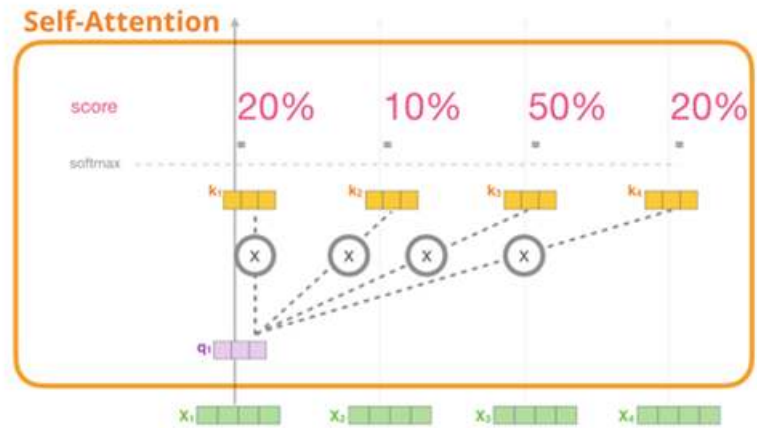


Figure 1.13: calculations of score *Alammar (2019)*.

## 2. Sum

We can now multiply the scores by the value vectors. A value with a high score will constitute a large portion of the resulting vector after we sum them up.

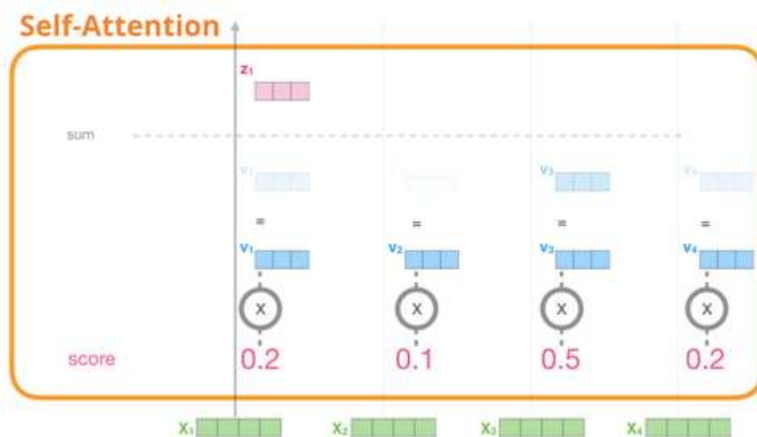


Figure 1.14: sum all vector to obtain  $Z$  *Alammar (2019)*.

### • The Illustrated Masked Self-Attention

We already knew that self attention would do the dot product query vector with all the keys in each word on the sentence input, but the masked self attention differs. When we try to know and predict the word after decoding process of the sentence, it is logical that we do not know the words that come after the word that we want to predict.

And this is exactly what the masked self attention does. It hides the words after the word to be predicted by multiplying its values by zero:



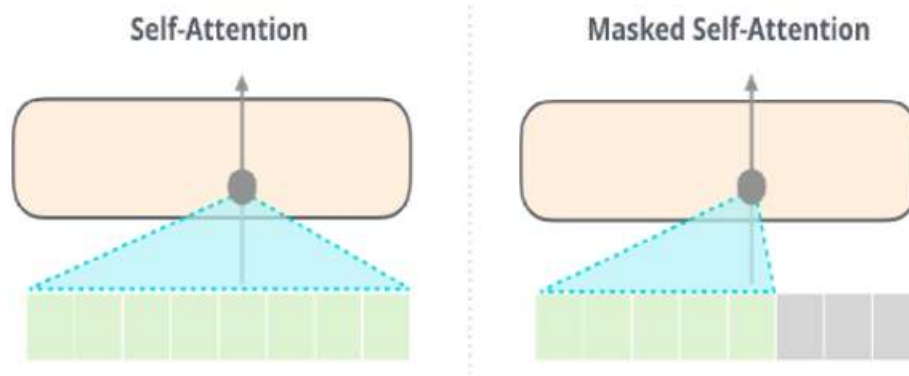


Figure 1.15: masked self and self attention [Alammar \(2019\)](#).

### • Multi-Head Attention

A sets of vectors, queries, keys and values used to expand the ability of attention, focus on different positions in the sentence input, the difference with using it is [Vaswani et al. \(2017\)](#).

If we calculate and get the Z vector using it we will find that it contains a better result but if we did not use it the result will always be dominated by the same word in each calculation [Zhang et al. \(2019\)](#).

In GPT-2 used 12 layer of self attention each layer has the matrix  $W_k$   $W_v$   $W_q$  :

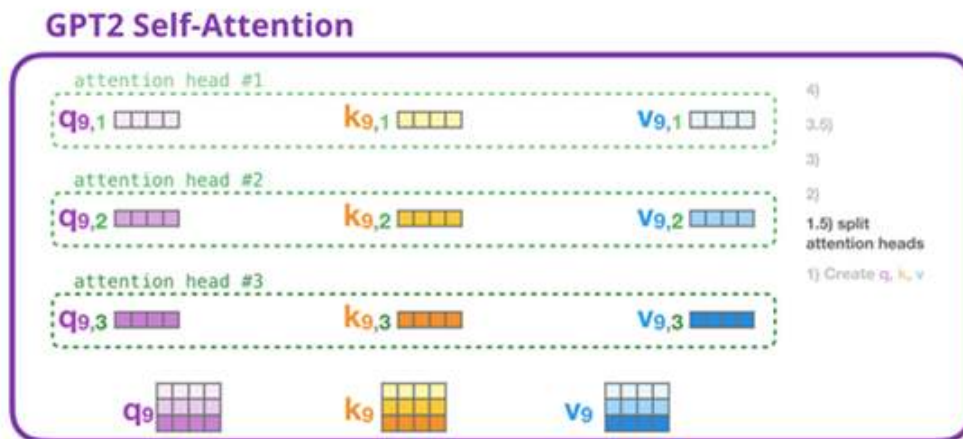


Figure 1.16: Multi-Head Attention -12 layers [Alammar \(2019\)](#)

### – Merging all the 12 multi-head In one vector

After getting the result of one to- ken embedding sum Z in each layer of multi-head We merge the result as the figure show



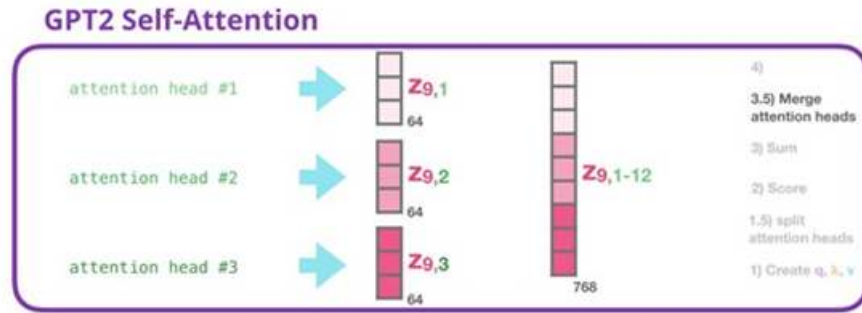


Figure 1.17: Merging all Heads.

– **Projection the merged vector with own matrix**

After getting all concatenated heads ( $z_1 \dots z_{12}$ ), it is produced with matrix  $Z$  to obtain the result vector that is able to go through the feed forward neural network.

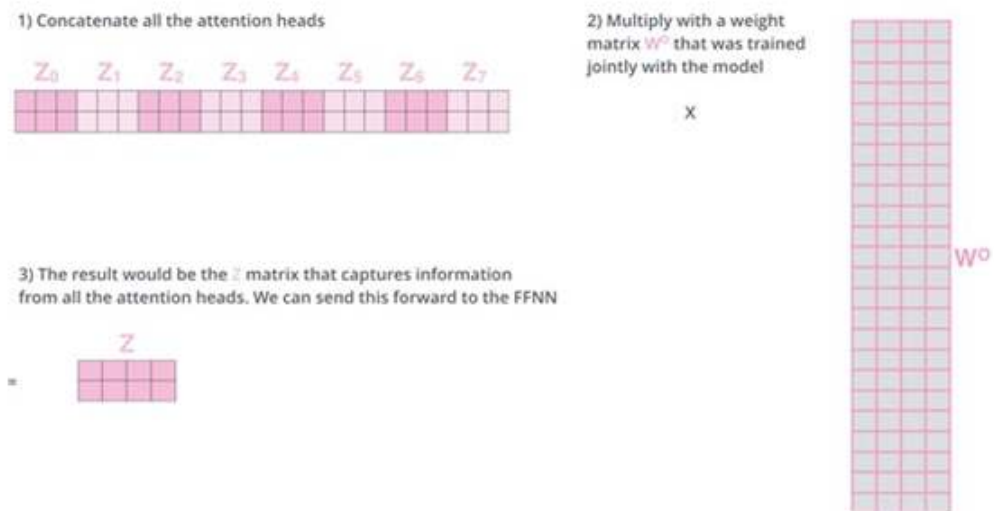


Figure 1.18: result of dot product between  $Z$  and all head.

- **GPT-2 Fully-Connected Neural Network Layer 1** After the dot product between the vector and the attn/c-proj/w matrix the output matrix will be sent to the feed-forward neural network. GPT-2 Fully-Connected is where the block processes its input token after self-attention has included the appropriate context in its representation. It is made up of two layers.

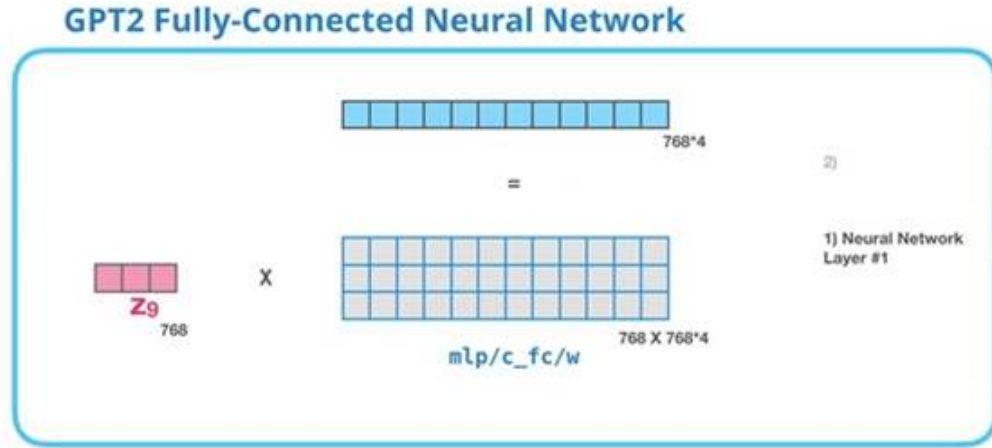


Figure 1.19: layer 1 [Alammar \(2019\)](#).

- GPT-2 Fully-Connected Neural Network layer 2** Layer 2 - Projecting to model dimension: The second layer projects the result from the first layer back into model dimension. The result of this multiplication is the result of the transformer block for this token.

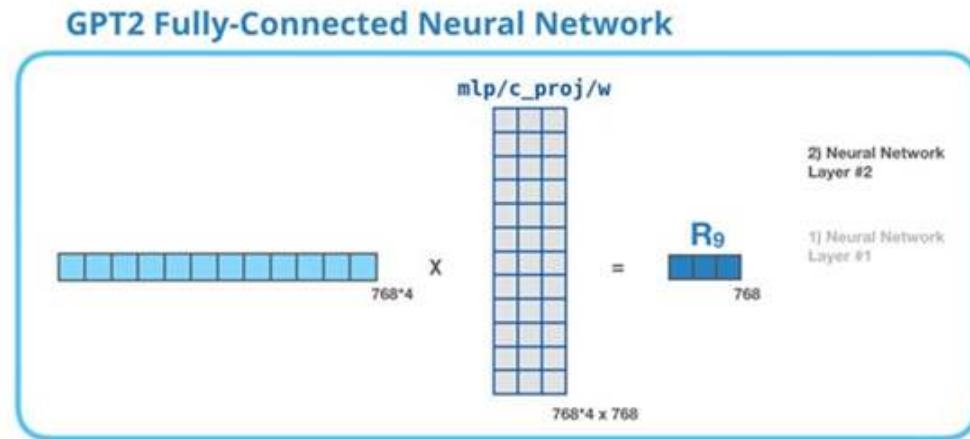


Figure 1.20: layer 2 [Alammar \(2019\)](#).

## 1.4.4 Grover Model

### 1.4.4.1 Definition

The positive of getting sophisticated techniques in NLP help the world to proceed but there is a negative in that, so creating fake texts became easier than before and in certain seconds [Zellers et al. \(2019\)](#).

The University of Washington professor said “Our work on Grover demonstrates that the best models for detecting disinformation are the best models at generating it”.

So Grover is a model that can generate a fake text and detects its generated outputs by accuracy of 92% [Zellers et al. \(2019\)](#).

#### 1.4.4.2 General Overview

Grover Models can Generate a strong Fake News article, and not only the body of the article, but also, news source, publication date, author list and the title. The researchers announced that the ‘best models for generating neural fake news are also the best models at detecting it’. Grover Generate Fake News as an adversarial game:

#### 1.4.4.3 Adversary

This system works on Generating Fake News that seems to be reality to humans and the machine to fool it. The goal of making a strong system generator is to create a strong detector, so it can defend against fake text and stories generated by humans or machines [Zellers et al. \(2019\)](#).

#### 1.4.4.4 Verifier

After generating many stories and news the verified system will classify news stories as real or fake. This system will have access to unlimited real news stories and few Fake News stories from a specific adversary [Zellers et al. \(2019\)](#).

#### 1.4.4.5 Description

- **Architecture** Many Language Models are different in their architecture between each other, Grover in this situation has the same architecture of GPT-2 Model, they present 3 versions of Grover as the same in GPT-2 1.4.3 too [Zellers et al. \(2019\)](#).
  - **Smallest Model Grover-Base** has 12 layers and 124 million parameters.
  - **Grover-Large** has 24 layers and 355 million parameters [Zellers et al. \(2019\)](#).
  - **Largest Model Grover-Mega**, has 48 layers and 1.5 billion parameters [Zellers et al. \(2019\)](#).
- **Dataset** The collection of Real News has been gathered in many corpus in CommonCrawl, by utilizing the newspaper python library they extract the Meta-Data and the body from each article and they created a dataset containing 5000 news domains indexed by Google news [Zellers et al. \(2019\)](#).

## 1.5 Related Work

In this section we will talk about other systems related to our work in this thesis, those systems can classify and detect fake data

### 1.5.1 Lair Liar Pants On Fire

this is title of paper the owner made system for detecting Fake News by using text Classification and Deep Learning [Wang \(2017\)](#).

#### 1.5.1.1 The Dataset Used

The owners of this paper were talking about Fake News and how to fight it using Text Classification. They collected 12.8K true sentences for their data set which they called “LIAR data set”, and collected only 221 sentences from a site web named POLITIFACT.COM for helping anyone who wants to work about the same idea in a variety of domains [Wang \(2017\)](#).

#### 1.5.1.2 Deeper Inside The Model

the owner of this paper used convolution Neural Network for Text Classification . they used it for detecting Fake News. In addition to that they used LSTM Model to optimize the process of Classification [Wang \(2017\)](#).

- **Embedding Vector**

The system proposed in the paper starts to embed token input and transform it into a vector, each input word in the system will do the same until the final word.

The meta-data in the system is considered as an input string too and it will transform into a vector to optimize the classification of the convolution neural network [Wang \(2017\)](#).

- **Convolution Neural Network**

After input word and meta-data transform into a vector the next step is to send them into network layer-1. [Wang \(2017\)](#) [Zhang et al. \(2019\)](#)

- **Layer-1**

This layer contains one filter (a filter is a matrix of number). the vector of input will produce the filter row by row until the end of matrix input

The result of this product is a vector called “feature”. After this operation the feature vector will be sent to the next Layer called pooling layer.

- **Pooling Layer**

The result of dot product between inputs and filter is a vector called feature, layer pooling takes this vector and takes the biggest value in that vector and stores it until another value comes

- **Fully Connected Layer**

After all sentences in the dataset go through neural’s layers and layer pooling combines all max unit values in all sentences and meta-data it goes all in a fully connected layer where the neural network continues to train and predict the class of sentences [Zhang et al. \(2019\)](#).

- **LSTM**

Is a sort of neural network which is used to classify the data and predict the next data in data distribution.

LSTM’s main work is to clean the information (this information is important

or not). By using 3 gates LSTM can remember the important information Zhang et al. (2019).

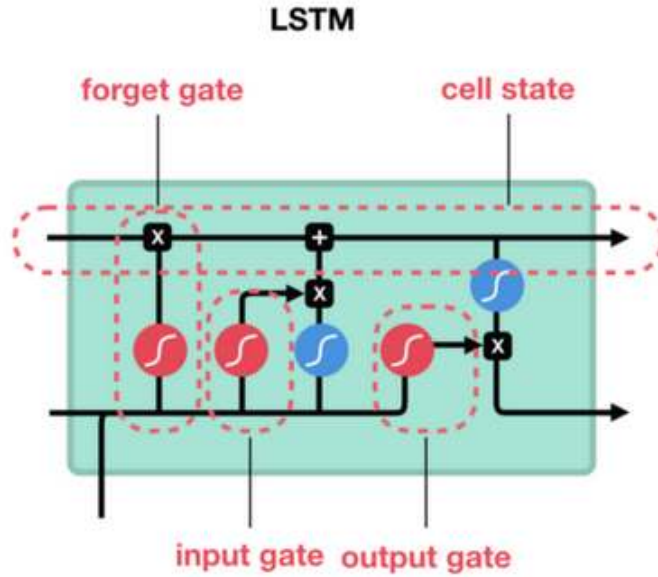


Figure 1.21: Architecture of LSTM Phi (2018)

\* Forget gate

It is considered as a neural network, it works by taking hidden state  $h_{t-1}$  and input  $x_t$  and combining them as one vector, then the vector goes as an input of that gate, the forget gate squashes the vector and makes it between 0 and 1 Zhang et al. (2019).

If the output of the forget gate is near to zero, it is considered as unimportant information, and if the output is near to one, it means the information is important. After getting the output of the forget gate, the result is stored in memory cell  $C$  Zhang et al. (2019).

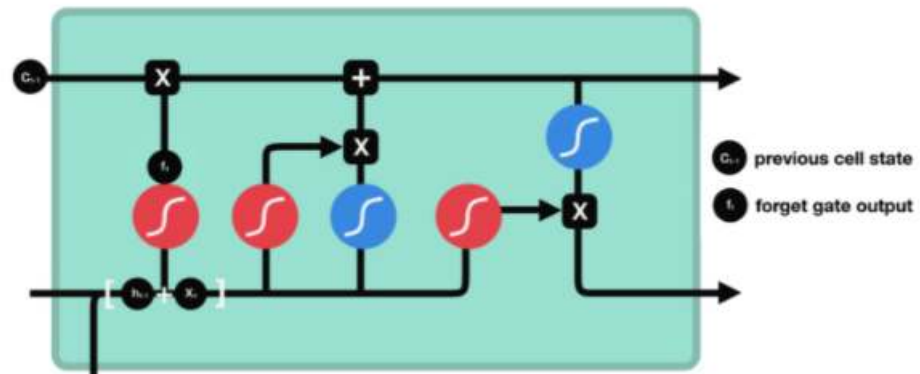


Figure 1.22: Forget gate and its output Phi (2018)

\* Input gate (update gate)

First, the hidden state and input  $x_t$  are combined as in the previous step, and this time it goes through two neural networks. The first is the input gate, which works the same as the previous gate, and the second uses the Tanh function to squish the output  $(h_{t-1} + x_t)$  between  $[-1, 1]$  Zhang et al. (2019).

After getting two outputs (output of sigmoid function and Tanh function), we will do a dot product between them, and the result goes to cell state  $C$ .

too.

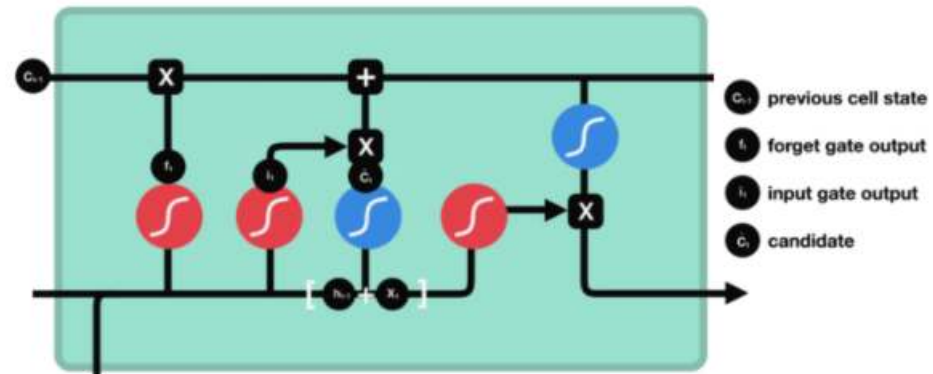


Figure 1.23: update gate architecture and it outputs [Phi \(2018\)](#)

\* How cell state update its information

After getting output in forget gate and getting output from two functions in input gate (update gate), the equation to calculus new cell state is:

New cell state = (output of forget function \* old cell state) + output of input gate [Zhang et al. \(2019\)](#).

\* Output gate

Finally the output gate takes input  $(h_{t-1} + x_t)$  and squishes it between 0 and 1, at the same time in the new cell stat  $C$  goes to Tanh function and transforms into a vector containing values between 1 and -1. The output of two inputs will produce with each other and the result is the new hidden state  $(h_t)$  in time step  $(t)$  [Zhang et al. \(2019\)](#).

### 1.5.1.3 Evaluation

Table 1.1: The evaluation results on the LIAR dataset. The top section: text-only models. The bottom: text + meta-data hybrid models Wang (2017).

Models	Valid.	Test
Majority	0.204	0.208
SVMs	0.258	0.255
Logistic Regression	0.257	0.247
Bi-LSTMs	0.223	0.233
CNNs	0.260	0.270
Hybrid CNNs		
Text + Subject	0.263	0.235
Text + Speaker	<b>0.277</b>	0.248
Text + Job	0.270	0.258
Text + State	0.246	0.256
Text + Party	0.259	0.248
Text + Context	0.251	0.243
Text + History	0.246	0.241
Text + All	0.247	<b>0.274</b>

## 1.5.2 Detecting Deceptive Reviews using Generative Adversarial Networks

Generative adversarial neural networks is an algorithms created to generate and classifier type of data, like: image, text, sound. . . ect . in generative adversarial network there are two main components The generator and discriminator. The Generator is a LSTM network that Generates Fake Data and the discriminator is a convolution Neural Network Aghakhani et al. (2018).

### 1.5.2.1 The Dataset Used

There are 3 type of dataset in this system Aghakhani et al. (2018)

- **Deceptive data:** contain about fake reviews.
- **True data:** contain about real reviews.
- **Generated data:** is the generator's outputs .

### 1.5.2.2 The architecture

The system are similar to a normal GANs but it differ on the old GANs in only with components, the system has 3 components one generator G and two discriminator D and D' [Aghakhani et al. \(2018\)](#) .

- **Discriminator D'**: used to detect between deceptive reviews and generated reviews.
- **Discriminator D**: used to distinguish between deceptive reviews and real reviews.
- **Generator G**: the generator model tries to generate fake reviews similar to D and D' not fake just to fool them.

### 1.5.3 Knowledge-Based Fake News Detection

To Detect Fake News in real life it often end to use method called Fact-checking, this method exploit the content of the news and compare it with others facts in the same domain of the news. Manual Fact-checking has two types the first is expert-based and the second is crowd-sourced fact-checking [Zhou and Zafarani \(2018\)](#).

#### 1.5.3.1 Expert-based Manual Fact-checking

The technique of expert based manual fact checking is based on get a lot of expert and extract the news from various sources and references and make compare manually. This technique are hard and take the time to implement [Zhou and Zafarani \(2018\)](#).

#### 1.5.3.2 crowd-sourced fact-checking

This method did not exploit on the expert to classify the news but the roller in this method is the crowd. The crowd is a set of people could entre to a various website and rate the article. This method is still in developing [Zhou and Zafarani \(2018\)](#).

#### 1.5.3.3 Automatic news fact checking process

In this process we have two phase the first phase is fact extraction and in this phase the data collected from web. after collecting data it come to next step the raw fact and in this step the process of cleaning fact are done and remove redundancy and invalidity and incompleteness data. After cleaning and prepare the data the next phase start.

In second phase there is fact-checking, it takes the article and compare the knowledge extracted from previous stage [Zhou and Zafarani \(2018\)](#).

## Conclusion

In this chapter we defined the concept of Language Model and some task of it like GPT-2 [1.4.3](#), Bert [1.4.2](#), and Grover [1.4.4](#) Models . In the same time we focus on GPT-2 model [1.4.3](#) and its component, what type and the dataset it was trained on. also we explained its functionality and how it works, moreover we mentioned The Related Work [1.5](#) of Detect Fake News.



## Chapter 2

# Collecting And Generating Datasat

<sup>1</sup> *"Without big data analytics, companies are blind and deaf, wandering out onto the web like deer on a freeway."*

*Geoffrey Moore*

## Introduction

The astonishing development of artificial intelligence is not to be underestimated especially in our days [Russell and Norvig \(2002\)](#), many people use the artificial intelligence in a positive stuff and finish their daily work by it, but there are many people use it in a negative purposes such as using their machines to generate disinformation and fake news on social media like: Facebook, Twitter, etc.especially in our Arabic society.

According to all of this result and problem there is no way but to fight against the fake news it must exist with a good amount of data to get good training.The problem suffers from a lack of Arabic news.In this chapter we will talk about steps of collecting our Arabic dataset and what tools we use to achieve that and we will talk about the statistics and analysis data set.In addition to all that we will use one of the previous language model for generates samples and building our Annotation for showing the results.

## 2.1 The Steps Of Collecting Data

For collecting our dataset we use python programming language, the library used in this process is called ‘ beautiful soup’.this library allows us to pull the article from any newspaper that we want and put them all in a certain location on our computer.

The beautiful soup library pulls the data by exploiting the architecture of sites (HTML code) [Richardson \(2007\)](#).

### 2.1.1 Collecting List of Arabic Newspapers

In any crawling process it must prepare the list of site names.that is what happened in our process.We collected and created a long list containing the title of Arabic newspapers.

### 2.1.2 Filtering The Newspapers

In this process we will enter every newspaper and try to choose it according to accessibility for each newspaper.(sometimes we find the news but in other form in HTML code).

### 2.1.3 The Write Of Code

By using beautifulsoup that we were talking about above we see the architecture of the site that we want to apply crawling on it and write the code for pulling all the articles from it.

There are two information in the article

- The main information: it contains with
  - The title of the article.
  - The date of the article.
  - The link to the article.
  - The content of the article.
- The sub-information: it contains on

- The writer of the article.
- The summary of the article.

### 2.1.4 The Process Of Crawl

After implementing the code of crawling we run it and we gathered all the collected articles in one file.

The process of classify and clean the domains

We create csv file for every domain in our own and classify all the related news by this domain according to following data

1. Name of newspaper in English.
2. Name of newspaper in Arabic.
3. The country of newspapers.
4. The link to the newspaper.
5. The original title.
6. The title of the newspaper after cleaning it.
7. The summary.
8. The summary after clean.
9. The original content
10. The original content after cleaning.
11. The link to the news.
12. The writer of news.
13. The date of news.

## 2.2 The Problem Faced Us

- **Structure of websites :** the problem here is the various pages of website in HTML coding so the position of news will be different from page to other one.
- **The limited crawling :** some websites block collection of their data after a certain number
- **The interruption of internet :** during crawling process sometimes the connection of internet goes off.

## 2.3 Data Analysis

We collected 5,551,441 news from various websites and newspapers only to do a good job and this news is in 50 newspapers (as it shows table 2.2) and classified them according to 17 domain (as it shows table 2.1) and 18 countries (as it shows table 2.1)

Table 2.1: Country and domains.

Country	Saudi	algeria	Britain	Egypt	Morocco	Syria	Sudan	Kuwait	Yemen
حالة الطقس و البيئة	0	0	0	0	0	0	0	0	0
سياسة	3	1645	0	170355	15407	4854	17309	10	725
اقتصاد	25575	13847	37	204993	6471	0	18495	4737	2089
علوم و تكنولوجيا	6	396	47	65410	871	0	0	0	0
دين	0	1424	0	0	0	0	0	1861	1839
رياضة	37563	75564	5811	227001	9103	0	8873	0	3187
موضة	3	807	0	41059	427	0	0	0	0
صحة	3	572	0	99814	1574	0	0	0	0
المجتمع	11100	99829	0	85310	27143	0	0	2733	2699
ثقافة و فن	34072	5114	34	136315	13398	0	0	845	983
منوعات	35215	13652	47	416110	34696	980	501	40	1000
تربية و تعليم	0	92	0	0	878	0	0	0	0
تواصل اجتماعي	5272	28	0	32855	0	0	0	0	0
اخبار محلية	68689	1198	0	133158	2447	0	14548	7192	18811
تاريخ	0	391	0	0	0	0	0	0	0
اخبار العالم	82687	155153	16118	645132	16208	2601	20997	3856	4958
صحافة	129108	1889	1070	886231	8237	0	30436	2546	3065
	429296	371601	23164	3143743	136860	8435	111159	23820	39356

Country	UAE	Tunisia	USA	Lebanon	Palestine	Bahrain	Iraq	Jordan	Qatar
حالة الطقس و البيئة	0	768	0	0	805	24	0	0	0
سياسة	0	12503	0	0	5352	0	0	0	923
اقتصاد	491	17755	170	0	9485	45	0	0	0
علوم و تكنولوجيا	8	2397	0	239	1708	0	0	0	0
دين	0	844	0	0	805	16	0	0	0
رياضة	25	17013	0	0	1933	25	0	138337	0
موضة	0	8097	0	0	0	14	0	0	0
صحة	4855	1801	0	514	810	18	0	0	0
المجتمع	95	6245	16996	2712	805	14	0	3962	0
ثقافة و فن	2	23416	42	700	2969	25	0	54907	0
منوعات	92	76106	17869	15206	29118	88	74471	110283	0
تربية و تعليم	59	10969	0	0	804	0	0	1728	0
تواصل اجتماعي	0	318	5535	2946	805	0	0	0	0
اخبار محلية	227	44080	219	0	20000	3	0	130430	0
تاريخ	0	651	0	0	0	0	0	0	0
اخبار العالم	1246	53974	3574	23123	15149	137	29036	192916	546
صحافة	2438	10379	8738	624	7151	21	32278	0	0
	9538	287316	53143	46064	97699	430	135785	632563	1469

Table 2.2: Newspaper and Country.

Arabic Name	English Name	News number	Country	Arabic Name	English Name	News number	Country
الجزيرة نت	Aljazeera	34	Saudi	الأيام السورية نت	Ayyam syria net	2601	Syria
الحدث	Alhadath	393		صدى الشام نت	Sadaa lshaam net	5834	
انحاء الالكترونية	an7a	45293		السوداني	Alsudani news	4379	Sudan
أم القرى	Uqngovsa	17809		ألوان ديلي	Alwan daily	527	
جريدة الرياض	Alriyadh	365767		السودان اليوم	Alsudan alyoum	106253	
الجديد اليومي	Eljadid Elyawmi	2088	algeria	الوسط	Alwasat	23820	Kuwait
جريدة الشعب	Ech chaab	14823		الثورة نت	Althawrah	34903	Yemen
جريدة المساء	el-massa	3169		الشارع	Alsharaeenews	4453	
الشروق اونلاين	Echorouk online	301897		الأيام	Alayam	5093	UAE
الخبر	Elkhabar	49624		البيان	Al bayan	4445	
بي بي سي عربي	BBC	376	Britain	الجديدة التونسية	Aljarida	116397	Tunisia
ميديا إيست اونلاين	Middle East Online	22788		النصر أونلاين التونسية	Assarih	64701	
طريق الأخبار	Akhbar Way	76229	Egypt	جريدة المغرب التونسية	Lemaghreb	28444	
الاهالي المصرية	Alahalygate	31812		حقائق أونلاين	Hakaek online	37158	
اليوم	Elyom	7034		الشروق التونسية	Alchourouk	40616	USA
صوت الأمة	Soutalomma	133497		بيروت تايمز	Beirut times	2294	
اليوم السابع	yom 7	2863901		صدى الوطن	Sada alwatan	4296	
البيان	El byan	31270		وطن	Watanserb	46553	
اخبار المغرب	Khabar maroc	2263	Morocco	أخبار الأرض	Cedar news	46064	Lebanon
يا بلادي	Yabiladi	25753		عرب 48	arab48	32232	Palestine
الزئقة 20	Rue 20	5632		الصمود	al somood	14479	
جريدة البيضاوي	Albidaoui	17064		جريدة الأيام	al-ayyam-ps	50988	
اصدقاء المغرب	Assdae	17595		أخبار الخليج	Akhbar alkhaeej	430	Bahrain
الصباح	Assabah	68553		الزمان	Azzaman	135785	Iraq
الرأية	alraiah	1469	Qatar	النستور	Addustour	632563	Jordan

## 2.4 Data Statistics

After collecting and cleaning the data, we ran some statistics on the collected data

- Average number of words in each field in table 2.3

Table 2.3: Average number of words.

Domain	average number word in title	average number word in summary	average number word in text
حالة الطقس و البيئة	9.38	0	325.63
سياسة	10.38	0	204.93
اقتصاد	9.81	0.003	232.73
علوم و تكنولوجيا	9.63	0.02	212.62
دين	6.27	0	586.74
رياضة	8.96	0.09	191.89
موضة	7.32	0.002	233.8
صحة	8.87	0.0007	233.91
المجتمع	11.24	0.003	239.34
ثقافة و فن	8.74	0.004	257.9
منوعات	8.68	0.002	226.83
تربية و تعليم	7.81	0	255.39
تواصل اجتماعي	8.8	0	131.56
اخبار محلية	09.07	0	190.38
تاريخ	8.63	0	1373.92
اخبار العالم	9.99	0.1	207.91
صحافة	12.73	0.01	330.85

- The average number of characters in each field [2.4](#)

Table 2.4: Average number of characters.

Domain	average number character in title	average number character in summary	average number character in text
حالة الطقس و البيئة	63.12	0	1987.79
سياسة	153.22	0	1292.96
اقتصاد	131.09	0.18	1455.69
علوم و تكنولوجيا	144.55	0.1	1295.11
دين	43	0	3474.1
رياضة	191.89	0.53	1196.66
موضة	141.3	0.01	1441.93
صحة	150.59	0.004	1431.78
المجتمع	92.28	0.002	1459.31
ثقافة و فن	110.63	0.02	1567.92
منوعات	98.66	0.01	1402.02
تربية و تعليم	52.09	0	1591.74
تواصل اجتماعي	132.38	0	892
اخبار محلية	92.89	0	1187.31
تاريخ	56.29	0	7830.81
اخبار العالم	117.77	0.58	1286.89
صحافة	168.02	0.04	2033.39

- The size (MB) of the data is collected in a txt file for each domain separately and for all domains are grouped in table 2.5

Table 2.5: The size of the data.

Domain	CSV (MB)	TXT (MB)
حالة الطقس و البيئة	6	5.68
سياسة	621	552.25
اقتصاد	888	805.03
علوم و تكنولوجيا	187	168.02
دين	34	41.45
رياضة	1000	1100
موضة	144	131.89
صحة	317	288.57
المجتمع	767	689.29
ثقافة و فن	841	773.15
منوعات	2000	2060
تربية و تعليم	44	41.48
تواصل اجتماعي	91	79.03
اخبار محلية	1000	960.15
تاريخ	14	14.5
اخبار العالم	3000	2930
صحافة	4000	4070
Sum	14954	54125

## 2.5 Generate Dataset

In order to generate Dataset from the data that you have collected, we need to use GPT-2, previously trained in Arabic language, and this form is



## 2.5.1 Arabic Pre-Trained GPT-2

We implemented a fake dataset starting from using model gpt-2 which is pre-trained in Arabic language by “AKHOLI” with a size of corpus 950 MB that he got from Wikipedia dataset. The model is available in : [Press here](#).

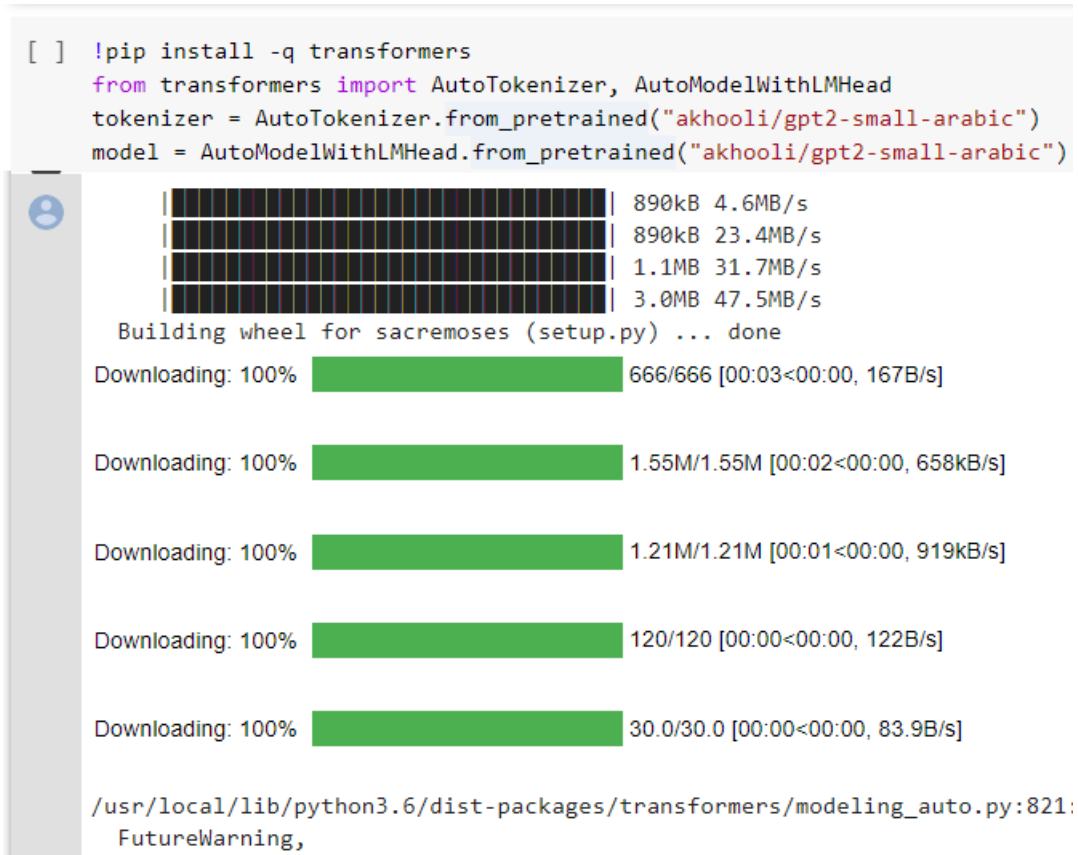


Figure 2.1: hugging face web page of gpt-2 small Arabic [AKHOLI \(2020\)](#)

### 2.5.1.1 Install Arabic Pre-Trained GPT-2

The architectures of gpt-2 is very complicated and to use it, must download and install Transformer and import AutoTokenizer, AutoModelWithLMHead.

```
[ ] !pip install -q transformers
from transformers import AutoTokenizer, AutoModelWithLMHead
tokenizer = AutoTokenizer.from_pretrained("akhooli/gpt2-small-arabic")
model = AutoModelWithLMHead.from_pretrained("akhooli/gpt2-small-arabic")
```



```
890kB 4.6MB/s
890kB 23.4MB/s
1.1MB 31.7MB/s
3.0MB 47.5MB/s
Building wheel for sacremoses (setup.py) ... done
Downloading: 100% 666/666 [00:03<00:00, 167B/s]
Downloading: 100% 1.55M/1.55M [00:02<00:00, 658kB/s]
Downloading: 100% 1.21M/1.21M [00:01<00:00, 919kB/s]
Downloading: 100% 120/120 [00:00<00:00, 122B/s]
Downloading: 100% 30.0/30.0 [00:00<00:00, 83.9B/s]

/usr/local/lib/python3.6/dist-packages/transformers/modeling_auto.py:821:
FutureWarning,
```

Figure 2.2: Install Arabic Pre-Trained GPT-2

### 2.5.1.2 Use Arabic Pre-Trained GPT-2

Picture figure 2.3 shows how to use the Arabic Pre-Trained GPT-2

```
def Generate_gpt (text):
    input_ids = tokenizer.encode(text, return_tensors='pt').to('cuda')
    model.to('cuda');
    max_length = 125
    sample_outputs = model.generate(input_ids, pad_token_id=50256,
                                   do_sample=True,
                                   max_length=max_length,
                                   min_length=500,
                                   top_k=40,
                                   num_return_sequences=1)

    s=""
    for i, sample_output in enumerate(sample_outputs):
        s+=str(tokenizer.decode(sample_output.tolist())).split(".")[0]
    return s
```

Figure 2.3: Code Arabic Pre-Trained GPT-2

- Fuction generate\_gpt use
  - input\_ids This variable express the value of the token
  - max\_length :it is an optional argument and it express on The maximum length of the sequence.

- `min_length` : it is an optional argument in the method `generate()`. It express on The minimum length of the sequence.
- `do_sample` : it is Boolean argument and if the value of it is **FALSE** that means greedy decoding is working else sampling is used.
- `Top-k`: its integer argument and it return to top k tokens

### 2.5.1.3 Example Arabic Pre-Trained GPT-2

figure 2.4 show example of some samples generated by Arabic Pre-Trained GPT-2

- هجوم مسلح على بنك في المحافظة الشمالية
- Sentence 0 :  
هجوم مسلح على بنك المغربسة العسكري الذي كان يعد العمود الفقري للدولة الإسلامية حتى عام 2010، حيث أطلق على هذا العمود اسم : 1. في 30 يونيو 2011، كشف محمد بن عبد الله أحد المقربين
- Sentence 1 :  
هجوم مسلح على بنك القاهرة المركزي بعد محاولة فاشلة لاغتيال رئيس حزب الحرية والعدالة آنذاك أحمد سالم صلاح، حيث تم توقيع اتفاقية مع منظمة التحرير الفلسطينية لإنهاء الحرب. في يناير 2014، بعد أيام من إطلاق نار
- Sentence 2 :  
هجوم مسلح على بنك الاحتياطي الإيراني في 15 سبتمبر 1990، بعد فشل محاولة اغتيال الزعيم الشيعي الخميني قبل فترة وجيزة من بداية الحملة العسكرية. وفي 15 سبتمبر 1990، أمر الجيش الإيراني بإطلاق سراح جميع السجناء
- Sentence 3 :  
هجوم مسلح على بنك مصر المصري في ذلك اليوم بعد أن ضاق أحد البنوك المالية على رئيس الوزراء المصري محمد حسني مبارك بتهمة الفساد، وقد أصدر مجلس الوزراء قرار بحظر بيع المشروبات الكحولية. وفي 22 مارس
- Sentence 4 :  
هجوم مسلح على بنك السودان المركزي والذي كان قد فقد في السابق وكان قد قام بانقلاب عسكري على حكم الشركة والتي أطلق عليها الرصاص الحي مما أدى إلى انهيار الحكم إلى أن تم تحرير معظم البلاد من حكومة
- Sentence 5 :  
هجوم مسلح على بنك الكويت الدولي في 15 نوفمبر 2011. في 26 نوفمبر 2011، اعتقل أحمد الجابر الصباح في الكويت لمدة ثلاث سنوات بتهمة تفجير انتحاري. في 22 ديسمبر / كانون الثاني، تعرض لمحاولة اعتداء
- Sentence 6 :  
هجوم مسلح على بنك إسرائيل الإسرائيلي، وتم استخدام نظام الإنذار السريع، مما أدى إلى انهيار النظام. لقد تم استخدام نظام الإنذار السريع، الذي كان يتم استخدامه في السابق في الهجمات المضادة على الفلسطينيين منذ عام
- Sentence 7 :  
هجوم مسلح على بنك بنجوين وهو بنك غير رسمي غير ربحي في عام 1995. في وقت لاحق من ذلك العام، تم تأسيس بنك بنجوين الجديد، وكان قد تم اختياره من قبل
- Sentence 8 :  
هجوم مسلح على بنك الرياض المركزي، والذي شارك في تفجير السفارة العراقية في الرياض في 1 أغسطس 2007. ومع ذلك، فإن الهجوم وقع في حوالي الساعة 3:00 مساء بتوقيت مكة المكرمة على الساعة السابعة
- Sentence 9 :  
هجوم مسلح على بنك الإسكان العسكري الصهيوني وكان قد اعتقل ما يقرب من ثلاثين شخصا خلال العملية وقد اتهمهم بالتغريد والنفي في السجون الإسرائيلية، بالإضافة إلى أن المحكمة اليهودية قد رفضت الادعاء. وتم

Figure 2.4: An example of generating sentences using Arabic Pre-Trained GPT-2

## 2.6 The Problems of Arabic Pre-Trained GPT-2

after using khooli's model we found some problems such as :

- **the replication of the words** : one of the most important problems in GPT-2 in general whether is trained on English or any other language is the replication of the words. This is due to several reasons like : the lack of the size in the dataset, lack of training time.
- **The understandable sentences** : sometimes the model generates understandable sentences by concatenating two different concepts that have no relationship with each other in one sentence.
- **The incomplete sentences** : the model generate good sentences but sometimes doesn't, this problems appears when the length of a good sentence doesn't fit with the number of a generated word

## 2.7 Solve The Arabic Pre-Trained GPT-2 Problems

In order to solve the aforementioned problems, we developed a algorithm for cleaning that complements its role in improving inputs and outputs

### 2.7.1 Arabic Pre-Trained GPT-2 ++

We implement the same previous steps that we describe above for generating 300 sentences, and by entering it in the algorithm for cleaning that we add it in gpt-2. the added algorithm for cleaning will optimize the generation of pre-trained models.

to optimize the result of gpt-2 Arabic pre-trained we remove duplication by using `remov_duplicates()`.

```

def remov_duplicates(input):

    # split input string separated by space
    input = input.split(" ")

    # joins two adjacent elements in iterable way
    for i in range(0, len(input)):
        input[i] = "".join(input[i])

    # now create dictionary using counter method
    # which will have strings as key and their
    # frequencies as value
    UniqW = Counter(input)

    # joins two adjacent elements in iterable way
    s = " ".join(UniqW.keys())
    return s

List = ['مباشرة', 'مباشرة', 'مباشرة']
import glob
import csv
for i in List :
    lm=[]
    Listed=[]
    index_1 = "/content/drive/My Drive/fake News 2 Master 2
    with open(index_1, 'rt') as fl:
        let1 = csv.reader(fl)

```

Figure 2.5: Code this algorithm for cleaning

## 2.8 Test Arabic Pre-Trained GPT-2 and Arabic Pre-Trained GPT-2 ++

After adding algorithm for cleaning the output of Arabic Pre-Trained GPT-2 we got Arabic Pre-Trained GPT-2++. In our plan we need to evaluate between Arabic Pre-Trained GPT-2 and Arabic Pre-Trained GPT-2++ so that is why we built the questionnaire. This questionnaire is for the humans and we will test the response of their own. After getting the response from the samples of humans we will compare it with Arabic Pre-Trained GPT-2 and Arabic Pre-Trained GPT-2++ result to know the power of the filtrate algorithm. And to compare between the result we use logical comparison.

### 2.8.1 Questionnaire

We selected 3 domains from the collected dataset that we crawled from various newspapers and randomly selected 300 sentences (titles) from each domain.

By using Arabic Pre-Trained GPT-2 We created 100 sentences starting from the first of four words in all 300 randomly selected sentences in each domain (300 sentences generated).

We take 100 sentences from each domain from the 300 randomly selected sentences and combine them with the 100 sentences generated for each domain. Now the number has become 600, which we mix and divide into six files (each file contains 100 sentences).

We repeat this process, but in this time with using Arabic Pre-Trained GPT-2 ++.

when the process of generation is done, we will receive 12 Annotations, for each one consist of 100 sentences. We hired 6 trusted people and assigned them to fill the Annotations

## 2.8.2 The Result

These are the results we got. The results represent the percentage of Fake sentences that were detected by people (for more details see the following table 2.6).

Table 2.6: The Result

Domain	Arabic pre-trained GPT-2		Arabic pre-trained GPT-2 ++		المجال
	model resolution	number sentences	model resolution	number sentences	
health	51 %	100	34 %	100	صحة
sport	49 %	100	35 %	100	رياضة
politic	51 %	100	38 %	100	سياسة
Total percentage	50.33 %	300	35.66 %	300	اجمالي النسبة المئوية

## 2.8.3 Examples

figure 2.6 show example of some samples generated in good syntax by Arabic Pre-Trained GPT-2 ++ and another samples generated in bad syntax by Arabic Pre-Trained GPT-2 . model resllution : (Fake And Real = 0 , Fake And Fake = 1 , Real And Real = 0 , Real And Fake = 0 ) .

GPT-2 ++	وزير الشباب: تكثيف الحوارات مع السياسيين الشباب
GPT-2	وزير الشباب: تكثيف الحوارات : نشروا كتابا بعنوان أنا لست أقيت لكم
-----	
GPT-2 ++	التوازن النفسي السليم يساعد بدوره على تقليل اضطرابات الاكتئاب
GPT-2	التوازن النفسي السليم يساعد على التخلص منه لأنه قد يضر بشكل خاص المشاكل النفسية
-----	
GPT-2 ++	س وج كل ما تم إنتاجه من مواد عضوية
GPT-2	س وج كل ما عدا ذلك فإن مثل هذه الأنواع لديها قدرة فائقة على إنتاج أكثر من أي نوع من الكائنات الحية الدقيقة، مما يسمح له بالاحتفاظ بأي نوع من الكائنات الحية الدقيقة
-----	
GPT-2 ++	دراسة تطالب بوقف النيورونتين إلى أن يصل عدد سكانها أكثر من 5 مليون نسمة
GPT-2	دراسة تطالب بوقف النيورونتين، أو السماح لها بدخول العالم الخارجي في مرحلة خروج المغلوب
-----	
GPT-2 ++	فواكه وخضراوات غنية بالفلافونويد تمنع السرطان في المخ
GPT-2	فواكه وخضراوات غنية بالفلافونويد تمنع السرطان من التنام الجروح لدى الإنسان، كما أن له دور في منع الأمراض الجلدية والرناسية
-----	
GPT-2 ++	أهم طرق علاج غثيان وقيء
GPT-2	أهم طرق علاج غثيان، ونزف، وتقيؤ، وغثيان، وأزيز، والتهاب شعبي، وتقيؤ غير مألوف من جميع أنحاء العالم
-----	
GPT-2 ++	ألوان ملكية وأعلام ترفرف في أنحاء العالم
GPT-2	ألوان ملكية وأعلام ترفرف ، على أن ينتقي إلى ملك جديد ويطلع على الملك الجديد
-----	

Figure 2.6: Example superiority of Arabic Pre-Trained GPT-2 ++ of Arabic Pre-Trained GPT-2

## Conclusion

In this chapter we talk about the lack of arabic dataset that contain fake and true news and introduced that in the introduction of the chapter.

Also we define the steps of collecting data and the main library that we use for crawling the data, and talk about collecting a list of newspapers. We also go on cleaning and classify our dataset in certain domains. after that we talk about and used GPT-2 model and what we apply on it to be better. In addition to that we build a Annotation and its results.

## Chapter 3

# Fake News Detector System

*"In this era of fake news and paid news artificial intelligence is more and more used as a political tool to manipulate and dictate common people, through big data, biometric data, and AI analysis of online profiles and behaviors in social media and smart phones. But the days are not far when AI will also control the politicians and the media too."*

*Amit Ray*



## Introduction

The problem in Natural Language Processing (NLP) [Manning and Schutze \(1999\)](#) is how to get the best representation of word, In chapter 1 we talked and define Word Embedding [1.3](#), Bag of Word (BOW) [1.3.4](#), represent the features of every method and also Related Work [1.5](#).

In this chapter we will see the implementation of the java system using a bag of words and deep learning that can classify the news that we crawled it and talk about it in previous chapter, in the same time we will use pre-trained Bert on our dataset and we will compare between them.

### 3.1 CBOW Based On Java System

Our system created in java programming language, it relies on Continuous Bag Of Word (COWB).Our system we utilize file to store and upload our data sentences.First step is cleaning the data in this step we took our file and remove every punctuation and all the symbols.In step 2 we extract all the sentences form dataset before cleaning it and put them all in file called structure dataset.

Our system contains 8 Layers each Layers has 2 type, first type is layer having activation function and in this layer we will not do a dot product between weights and inputs.The second type of our layers in neural network have Linear function and the different between the first type and the second type is the second type is using Leaner function for calculation weights to next Layer.

In Nearal Network there are two phase, the first phase is the training phase, in it we will take all lines in our dataset line by line and entering it in neural network ,and calculate the weights for each layers in forward and backward propagation and store them all in array list. After forward and backward propagation and after the storing of our vectors we will sum them until the batch size reach to 50 (batch size is variable in neural network).The testing phase comes after and in this phase we will test the accuracy of neural network.

#### 3.1.1 Training information of CBOW Based On Java System

We trained the model on Dataset containing 10,412 sentences (as it shows table [3.1](#)) and one cycle due to the lack of robust devices for model training.

*Table 3.1: Training information of CBOW Based On Java System*

epochs	Number of sentences	Accuracy
1	10412	47%

#### 3.1.2 Test Phase

In test phase we have 2000 sentences and we entered every line on our test data to our java detector and extract the accuracy.

### 3.1.3 The Class Of Java Project

is an explanation of all classes and functions CBOW Based On Java System [3.1](#)

#### 3.1.3.1 Class Layer

The following functions is a pseudo code of the main fuction of class Layer.

- Method Initweights

```
initWeights(int nextLayerSize) {  
    Random rn = new Random();  
    double nls = nextLayerSize;  
    double sd = 1 / Math.sqrt(nls);  
    weights = new double[neuroNumb][nextLayerSize];  
    for (int i = 0; i < weights.length; i++)  
    {  
        for (int j = 0; j < weights[0].length; j++)  
        {  
            weights[i][j] = rn.nextGaussian() * sd;  
        }  
    }  
}
```

*Figure 3.1: Method Initweights*

- ActFunc

```
public double[] activate()
{
    double[] act = new double[neuroNumb];
    if (isLinear())
    {
        act = MatrixOper.mulWithVec(MatrixOper.transpose(weights), vals);
    }
    else // activation function layer (ReLU for example)
    {
        for (int i = 0; i < act.length; i++)
        {
            act[i] = actFunc.act(vals[i]);
        }
    }
}
```

*Figure 3.2: ActFunc*

- CalcGrads .

```

public double[] calcGrad(double[][] prevLayWeights, double[] prevActDeriv)
{
    double[] grd = new double[neuroNumb];
    if (!isLinear())
    {
        grd = MatrixOper.mulWithVec(prevLayWeights, grad);
    }
    else // a linear layer
    {
        for (int i = 0; i < grd.length; i++)
        {
            //prevActDeriv is the derevatives
            grd[i] = grad[i] * prevActDeriv[i];
        }
    }
}

```

Figure 3.3: CalcGrads

- ElementWiseDrive

```

public double[] elementWiseDeriv()
{
    if (!isLinear())
    {
        double[] ewd = new double[vals.length];
        for (int i = 0; i < ewd.length; i++)
        {
            ewd[i] = actFunc.dAct(vals[i]);
        }
        return ewd;
    }
    return null;
}

```

Figure 3.4: ElementWiseDrive

### 3.1.3.2 Class NeuralNetwork

The following code are the pseudo code of main function of this class.

- InitStochGradDescn

```
private void initStochGradDesc()
{
    stochGradDesc = new ArrayList<>(); // same size as "lays" attribute
    for (Layer lay : lays)
    {
        stochGradDesc.add(new double[lay.getNeuroNumb()]);
    }
}
```

Figure 3.5: *InitStochGradDescn*

- initAccumulVals

```
private void initAccumulVals()
{
    accumulVals = new ArrayList<>(); // same size as "lays" attribute
    for (Layer lay : lays)
    {
        accumulVals.add(new double[lay.getNeuroNumb()]);
    }
}
```

Figure 3.6: *initAccumulVals*

- LossFunction

```
private double loss(double obsrvdVal, double calcVal)
{
    return Math.log(1 + Math.exp(-obsrvdVal * calcVal));
}
```

Figure 3.7: *LossFunction*

- DLossFunc

```
private double dLossForOutput(double obsrvdVal, double calcVal)
{
    return (-calcVal * Math.exp(-obsrvdVal * calcVal)) / (1 + Math.exp(-obsrvdVal * calcVal));
}
```

Figure 3.8: *DLossFunc*

- UpdatValsOf

```
private void updateValsOf(int layIndx) throws Exception
{
    if (layIndx == 0) { throw new Exception("index 0 is for input layer. It is
not updatable"); }
    lays.get(layIndx).setVals(lays.get(layIndx - 1).activate());
}
```

Figure 3.9: UpdatValsOf

- UpdatGradsOf

```
private void updateGradsOf(int layIndx) throws Exception
{
    if (layIndx == lays.size() - 1) { throw new Exception("index
"+layIndx+" is for output layer. It is not backpropa"); }

    lays.get(layIndx).setGrad(lays.get(layIndx + 1).calcGrad(
        lays.get(layIndx).getWeights(),
        lays.get(layIndx).elementWiseDeriv()));
}
```

Figure 3.10: UpdatGradsOf

- trinWithDSL

```
public void trainWithDSLLine(double[] bow, double label)
{
    lays.get(0).setVals(bow);
    for (int i = 1; i < lays.size(); i++)
    {
        updateValsOf(i);
    }
    double dloss = dLossForOutput(label, lays.get(lays.size() -
1).getVals()[0]);
    lays.get(lays.size() - 1).setGrad(new double[] {dloss});

    for (int i = lays.size() - 2; i >= 0; i--)
    {
        updateGradsOf(i);
    }
}
```

Figure 3.11: trinWithDSL

- train

```

train(String dsFile)
{
    int lnum = 0; //the number of lines.
    LexiconExtractor.initLexiconFromFile();
    initAllWeights();

    for (int epo = 0; epo < epochs; epo++)
    {
        br = new BufferedReader()
        int batchOffset = 0
        String ln = br.readLine();

DO
        lnum++;

        String labS = ln.charAt(ln.length() - 1) + "";

        double label = Double.parseDouble(labS);
        double[] bow = extractBoW(ln);
        trainWithDSLLine(bow, label);
        for (int i = 0; i < lays.size(); i++)
        {
            Acumul();
            StochastGrads

        }

        batchOffset++;

        if (batchOffset == batchSize)
        {
            {
                updateAllWeights();
                batchOffset = 0;
                initStochGradDesc();
                initAccumulVals();
            }
            System.out.println("a line :"+lnum);
            ln = br.readLine();
        }

while(ln != null);

        br.close();
        if (batchOffset != 0)
        {
            updateAllWeights();
            initStochGradDesc();
            initAccumulVals();
        }
        storeWeightsInFile();
    } // end of epochs
}

```

Figure 3.12: train

### 3.1.4 CBOW Based On Java System Results

The following table 3.2 shows the results of applying the Questionnaire (2.8.1) to the form

Table 3.2: CBOW Based On Java System Results

Domain	Arabic pre-trained GPT-2		Arabic pre-trained GPT-2 ++		المجال
	model resolution	number sentences	model resolution	number sentences	
health	64 %	100	65 %	100	صحة
sport	55 %	100	55 %	100	رياضة
politic	65 %	100	65 %	100	سياسة
Total percentage	61.33 %	300	61.66 %	300	اجمالي النسبة المئوية

#### 3.1.4.1 CBOW Based On Java System Results VS Person

- Fake Sentence, output CBOW Based On Java System Fake, output Person Fake :

الجملة رقم 1 :

فرصة مصر سيناريوهات مختلفة للتأهل إلى أقصى حد

الجملة رقم 2 :

صور على عبد العال، وقد توفي في دمشق ودفن هناك

الجملة رقم 3:

منتخبنا يواجه اليابان ويسعون إلى تحقيق بعض النجاحات، وهذا من جهة أن الجميع يعتبرهم أحد أعظم المرشحين، لأن هناك العديد من المواطنين اليابانيين الذين لم يجدوا فرص في التصويت بسبب مشاركتهم انتخابات عام 1999

Figure 3.13: Fake Sentence, output CBOW Based On Java System Fake, output Person Fake

- Fake Sentence, output CBOW Based On Java System Real, output Person Fake :



**الجملة رقم 1:**

دیل بوسکی علی اتصال هاتفي مع شركة وومانز التي اشترت حقوق الملكية الفكرية، ولكن لم تكن قد كانت جزءا منها

**الجملة رقم 2:**

الأهلي يغادر بتسوانا بعد أن تركته البلاد

Figure 3.14: Fake Sentence, output CBOW Based On Java System Real, output Person Fake

- Fake Sentence, output CBOW Based On Java System Real, output Person Real:

**الجملة رقم 1:**

إنفوجراف مانشستر يونايتد لا زال مستمرا في الدوري منذ 2005 و 2007

**الجملة رقم 2:**

نصائح لصاحبات الشعر الخفيف والأقدام الطويلة

**الجملة رقم 3:**

4 نصائح للتغلب على المشكلات التي يعاني منها

**الجملة رقم 4:**

فيديو معلوماتي فواكه هتعاظ على طعم الأرض ولونها الأخضر

**الجملة رقم 5:**

4 أمراض تصيب المعدة والأمعاء

Figure 3.15: Fake Sentence, output CBOW Based On Java System Real, output Person Real

- Fake Sentence, output CBOW Based On Java System Fake, output Person Real:

- الجملة رقم1:**  
ملخص وأهداف مباراة توتنهام هوتسبير
- الجملة رقم2:**  
لماذا يرفض الأهلي تكريم الملك فيصل
- الجملة رقم3:**  
النيابات العامة المتخصصة في العلاج
- الجملة رقم4:**  
فيتامين ك يقلل من مستويات هرمون التوتنويد لدى النساء

Figure 3.16: Fake Sentence, output CBOW Based On Java System Fake, output Person Real

## 3.2 Detector System Based On Bert

As we talk in chapter one Bert is a bidirectional system used in many domains like text generation, text classification.

The Bert system is different from our java system, it uses word embedding to transform a word in vector and transformers block in addition to the neural network for prediction.

This model is similar to GPT-2 in transformer architectures and both of them use attention mechanisms to get the context between the tokens.

### 3.2.1 Training information of Detector System Based On Bert

On two databases, the first one consists of 11 and the second of We trained the model on two Dataset, the first one consists of 210,000 sentences the second of 80,000 sentences (as it shows table 3.3) and one cycle due to the lack of robust devices for model training.

Table 3.3: Training information of Detector System Based On Bert

Name Detector Bert Systems	Number of sentences	epochs	Accuracy
BERT 210,000 sentence 10 epochs	210,000	10	99.95%
BERT 80,000 sentence 10 epochs	80,000	10	99.94%

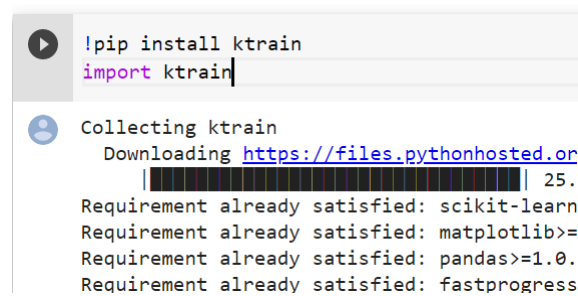
## 3.2.2 The implementation of Bert system for classify the news

### 3.2.2.1 Ktrain Library

The system uses ktrain library it helps building and training the neural network, the library also makes the code of training be reduced in a few lines. Ktrain also allows us for :

- It uses the learning rate finder to estimate an optimal learning rate
- save model after training and loading it in any time you want.

To import ktrain write the following line.



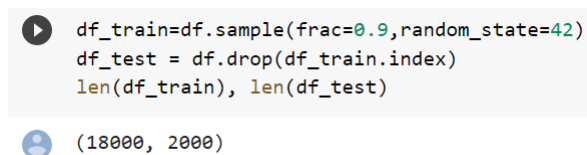
```
!pip install ktrain
import ktrain
```

Collecting ktrain  
 Downloading <https://files.pythonhosted.org/...> 25.  
 Requirement already satisfied: scikit-learn  
 Requirement already satisfied: matplotlib>=  
 Requirement already satisfied: pandas>=1.0.  
 Requirement already satisfied: fastprogress

Figure 3.17: Install and Import Ktrain

### 3.2.2.2 Splitting DataSet

The method sample works on splitting the data set by percent and the rest of the dataset is the second line of code. It means if the dataset has 1000 sentences and we got 90% to train and 10% for test it will be 900 for training and 100 for testing.



```
df_train=df.sample(frac=0.9, random_state=42)
df_test = df.drop(df_train.index)
len(df_train), len(df_test)
```

(18000, 2000)

Figure 3.18: Splitting DataSet

### 3.2.2.3 Loading And Using Pre-Train Model

To implement a news classifier we used pre-train Arabic model called 'arabert', and we train it using our dataset in 5 epochs. In the following figure we provide the model after loading it by training part and test part of our dataset and we launch it to training.

```

▶ from ktrain import text
MODEL_NAME = 'aubmindlab/bert-base-arabertv01'
t = text.Transformer(MODEL_NAME, maxlen=128)
trn = t.preprocess_train(df_train.review.values, df_train.rating.values)
val = t.preprocess_test(df_test.review.values, df_test.rating.values)
model = t.get_classifier()
learner = ktrain.get_learner(model, train_data=trn, val_data=val, batch_size=32)
learner.fit_onecycle(5e-5, 12)

preprocessing train...
language: ar
train sequence lengths:
  mean : 7

```

Figure 3.19: Loading And Using Pre-Train Model

### 3.2.2.4 Prepare To Predict And Save

After loading Ara-Bert and train it on large data set the next step is the prediction so to predict it we write the following code :

```

[ ] p = ktrain.get_predictor(learner.model, t)

[ ] p.save(path)

```

Figure 3.20: Prepare To Predict And Save

### 3.2.2.5 Load and Use our system to predict some examples

```

[ ] Sentence = 'البرسا تفوز الدوري الانجليزي في البرازيل'
num=40000
path = "/content/drive/My Drive/fake News 2 Master 2020/"
ktrain.load_predictor(path).predict(Sentence)

'Fake'

```

Figure 3.21: Load and Use our system to predict

## 3.2.3 Detector System Based On Bert Results

The following table 3.4 shows the results of applying the Questionnaire (2.8.1) to the form

Table 3.4: Detector System Based On Bert Results

Domain	BERT 210000 sentence 10 epochs		BERT 80000 sentence 10 epochs		المجال
	Arabic pre-trained GPT-2++	number sentences	Arabic pre-trained GPT-2++	number sentences	
health	96 %	100	96 %	100	صحة
sport	88 %	100	89 %	100	رياضة
politic	96 %	100	95 %	100	سياسة
Total percentage	92.66 %	300	93.33 %	300	اجمالي النسبة المئوية

### 3.2.3.1 Bert 210,000 Sentences 10 Epochs VS Person

- Fake Sentence, output Bert 210,000 Sentences 10 Epochs Fake, output Person Fake :

الجملة رقم 1:

مفاجأة انتحار 3 من الذكور و الإناث 1 النساء في نفس الفئة العمرية 9 أو أكثر 16 15 سنة

الجملة رقم 2: عاهد يقدم بتأزة محكيات أو سلع شخصية

الجملة رقم 3: الأهلي يغادر بتسوانا بعد أن تركته البلاد

الجملة رقم 4: الناصرة: استنكار التهجم على أي شيء، ولا أن تكون صحيحة

الجملة رقم 5 :

حدث اليوم الأولمبية تعتمد عليه أغلب الدول المشاركة فيها، بما فيها الإسكندنافية، كما أنها تشارك في أولمبياد عام 2010 ، وتعتبر هذه المجموعة من أهم الأولمبياد الشتوي، والتي تنظم لمدة ثلاثة أسابيع.

الجملة رقم 6 :

صريف الأسنان ليلاً دليل على أن السبب غير معروف في هذه النظرية

Figure 3.22: Fake Sentence, output Bert 210,000 Sentences 10 Epochs Fake, output Person Fake

- Fake Sentence, output Bert 210,000 Sentences 10 Epochs Real, output Person Fake :

**الجملة رقم 1:**

اتحاد الرمثا واليرموك شوقك يا جمال

**الجملة رقم 2:**

اعرف مقدار الفيتامينات الصحيحة والحالات التي لم يتم تحديد سبب  
هذه الحالة

**الجملة رقم 3:**

حمدى الفخرانى: بديع مختبئ يورى: أبو علي كرمى: عبد الله محمد  
مصطفى

Figure 3.23: Fake Sentence, output Bert 210,000 Sentences 10 Epochs Real, output Person Fake

- Fake Sentence, output Bert 210,000 Sentences 10 Epochs Real, output Person Real:

الجملة رقم 1 : ملخص وأهداف مباراة توتنهام هوتسبير

الجملة رقم 2: تدخين الأم وتعاطيها الكحوليات يعرض الطفل للبيع في المتاجر

الجملة رقم 3: رئيس المحافظين: الترشح للبرلمان فرض كفاية على مجلس الشيوخ مناقشة الدستور

الجملة رقم 4: عمرو طارق يكشف سبب مرض كوفيد 19

الجملة رقم 5: لماذا يرفض الأهلي تكريم الملك فيصل

الجملة رقم 6: بوبى تشارلتون: رونالدو يغير اسم النادي إلى الدوري الهولندي

الجملة رقم 7: تقارير: رابطة الدوري الألماني لكرة القدم

الجملة رقم 8: صحتك في وصفة شوربة من حبوب الشاي

الجملة رقم 9: الزمالك يرفع سقف رواتب الموظفين

الجملة رقم 10 : مظاهرات في ألمانيا للمطالبة بالامتثال للإجراء

Figure 3.24: Fake Sentence, output Bert 210,000 Sentences 10 Epochs Real, output Person Real

- Fake Sentence, output Bert 210,000 Sentences 10 Epochs Fake, output Person Real:

- الجملة 1: الإجهاض يدفع النساء للاكتئاب وعدم القدرة على استخدام العقاقير المضادة
- الجملة 2: مباراة كبيرة لليابان في 1 فبراير 1942، بعد أن قام هتلر بهجوم على جزيرة باراميسيا ضد بيرل هاربر
- الجملة 3: فوائد الزعتر لصحة الجسم مثل زيت الزيتون الذي يصنع من شجرة البند
- الجملة 4: سرطان: علاج تجريبي يثبت فعاليته في علاج الحالات المرضية
- الجملة 5: أنوميا الألوان تظهر لدى الأفراد الذين يعيشون في نفس المنطقة
- الجملة 6: الهرمونات التعويضية تسبب سرطان الثدي، حيث أن هرمون الألم عادة ما يكون مرتبطاً بتلف الأعضاء التناسلية
- الجملة 7: اكتشاف البروتين المتسبب في حدوث أي خلل في الجهاز التنفسي
- الجملة 8: أهم طرق علاج غثيان وقيء
- الجملة 9: شاهد نيمار يساهم في تشكيل العديد من الفعاليات الثقافية والاجتماعية والثقافية في لبنان والعالم العربي
- الجملة 10: وزيرة البيئة تؤكد أهمية هذه العلاقة بين العلم والصناعة، على الرغم من بعض تلك التحديات

Figure 3.25: Fake Sentence, output Bert 210,000 Sentences 10 Epochs Fake, output Person Real

### 3.2.3.2 Bert 80,000 Sentences 10 Epochs VS Person

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Person Fake :

- الجملة رقم 1 : قيادي بالمصري الديمقراطي يبرئ من خلال استغلاله لبعض الأفكار الخاطئة في المجتمع المصري المعاصر الذي يحاول ان يلهي يشاء أجل الوصول إلى حل سريع و هو هذا ليس إلا خدعة أجل تحقيق الهدف الأساسي لأدراكني عالم السياسة
- الجملة رقم 2: العلاجات الموجهة تنصدر أعمال التصميم الفرنسي المعماري، الذي كان في بداية الحرب الباردة قد توقف عن المشروع فرنسا
- الجملة رقم 3: اتحاد الكرة يبحث عن طرق جديدة لتحقيق أهداف اللعبة، بينما يريد الفريق الآخر وضع خطة لجعل اللعب أكثر أهمية، ولكنه يفضل لعب من ذلك، لأنه يجب عليه أن يلعب فقط مباراة واحدة، لكنه يطلب اللاعبين يلعبوا نفس البطولة على الأقل
- الجملة رقم 4: السيسي في الخرطوم الخميس الموافق 8 أكتوبر 2015 وبعد ساعات من انتهاء عملية نقل البضائع السودان قام فريق اتحاد النقل العام السودانية بإجلاء عمال الإغاثة منازلهم إلى العاصمة
- الجملة رقم 5: مفاجأة انتحار 3 من الذكور و الإناث 1 النساء في نفس الفئة العمرية 9 أو أكثر 15 16 سنة
- الجملة رقم 6: هل بدأ تغشى فيروس 1 2 في وقت لاحق، وفي ديسمبر عام 2005، أصبح الإيدز سببا اكتشاف الورم الحليمي البشري

Figure 3.26: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Person Fake

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Person Fake :

الجملة 1:  
اتحاد الرمثا واليرموك شوقك يا جمال

الجملة 2:  
حمدي الفخراني: بديع مختبئ يورى: أبو علي كرمي: عبد الله  
محمد مصطفى

Figure 3.27: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Person Fake

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Person Real:

الجملة رقم 1: ملخص وأهداف مباراة توتنهام هوتسبير

الجملة رقم 2: التجمع : موقف الإنقاذ و ما تم الإعلان عنه من أمر إلا بتهديد لا مفر منه

الجملة رقم 3: تدخين الأم وتعاطيها الكحوليات يعرض الطفل للبيع في المتاجر

الجملة رقم 4: رئيس المحافظين: الترشح للبرلمان فرض كفاية على مجلس الشيوخ  
مناقشة الدستور

الجملة رقم 5: كاجاوا أفضل لاعب في العالم

الجملة رقم 6: فواكه وخضراوات غنية بالفلافونويد تمنع السرطان في المخ

الجملة رقم 7: العلاج الدوائي أفضل من النفسي

الجملة رقم 8: عمرو طارق يكشف سبب مرض كوفيد 19

الجملة رقم 9: بوبى تشارلتون: رونالدو يغير اسم النادي إلى الدوري الهولندي

الجملة رقم 10 : دالاس على بعد فوزين

الجملة رقم 11 : دراسة حديثة تكشف عن وجود علاقة بين الأرض والسماء مع كوكب  
المريخ

Figure 3.28: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Person Real

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Person Real:



- الجملة 3: بارترا يقرر مصيره مع رفاقه ومن معه
- الجملة 4 : الجمعية المصرية للقانون الدولي هي منظمة دولية من منظمات المجتمع المدني، تأسست في أكتوبر عام 1948 ، وبدأت العمل خلال وزارة العدل أوائل 1964، وكانت ضمن الأعضاء المؤسسين لهذا المنظمة
- الجملة 5: ما أسباب تلوث البيئة: زيادة استخدام المياه الملوثة الناتجة عن هذه الغازات في المناطق الحضرية
- الجملة 6: 40 مليون يورو لتونالي من الشياطين الحمر أو أي شيء آخر
- الجملة 7: صفحة أندلخت تستقبل 100 مليون دولار أمريكي
- الجملة 8: تصفيات المونديال العراق في شهر يوليو لعام 2016، عندما تم استبعاد اثنين من اللاعبين الأربعة الرئيسيين خلال مباراة ودية ضد منتخب سوريا لكرة القدم في مدينة الإسكندرية نهائي كأس العربية 2014.
- الجملة 9: أجويرو يقود هجوم السيتي على جزيرة كورفو حيث يمكن أن يقوم بتوصيل الطائرة إلى المياه المحيطة
- الجملة 10: الاكتئاب ومشاكل الرؤية أبرز عوامل النجاح في هذا المجال

Figure 3.29: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Person Real

### 3.2.3.3 Bert 80,000 Sentences 10 Epochs VS Bert 210,000 Sentences 10 Epochs

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Bert 210,000 Sentences 10 Epochs Fake :

- الجملة رقم 1 بارترا يقرر مصيره مع رفاقه ومن معه
- الجملة رقم 2: خطوات عملية للتعامل مع الرسائل، التي غالبا ما تكون محدودة جدا
- الجملة رقم 3: صفحة أندلخت تستقبل 100 مليون دولار أمريكي
- الجملة رقم 4:
- إخوان بلا عنف: التنظيم الإرهابي المعروف باسم القاعدة الشعبية في ليبيا
- الجملة رقم 5: القلق والاكتئاب والوسواس والفوبيا والأفراح والحواس والخطر
- الجملة رقم 6: القلق والاكتئاب والوسواس والفوبيا والأفراح والحواس والخطر
- الجملة رقم 7:
- مفاجأة انتحار 3 من الذكور و الإناث 1 النساء في نفس الفئة العمرية 9 أو أكثر 15 سنة 16

Figure 3.30: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Bert 210,000 Sentences 10 Epochs Fake

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Bert 210,000 Sentences 10 Epochs Fake :

الجملة رقم 1 :  
 العلاج الدوائي أفضل من النفسي  
 الجملة رقم 2 :  
 دالاس على بعد فوزين  
 الجملة رقم 3:  
 عدلي منصور: مصر ستظل في حاجة إليها

Figure 3.31: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Bert 210,000 Sentences 10 Epochs Fake

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Bert 210,000 Sentences 10 Epochs Real:

الجملة رقم 1:  
 ملخص وأهداف مباراة توتنهام هوتسبير  
 الجملة رقم 2:  
 تدخين الأم وتعاطيها الكحوليات يعرض الطفل للبيع في المتاجر  
 الجملة رقم:  
 فواكه وخضراوات غنية بالفلافونويد تمنع السرطان في المخ  
 الجملة رقم 4:  
 عمرو طارق يكشف سبب مرض كوفيد 19

Figure 3.32: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Real, output Bert 210,000 Sentences 10 Epochs Real

- Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Bert 210,000 Sentences 10 Epochs Real:

**الجملة رقم 1:**  
فيديو معلوماتي فواكه هتحافظ على طعم الأرض ولونها الأخضر

**الجملة رقم 2:**  
لماذا يرفض الأهل تكريم الملك فيصل

**الجملة رقم 3:**  
الاستفادة من اللاعبين غير المحترفين

**الجملة رقم 4:**  
الزمالك يرفع سقف رواتب الموظفين

*Figure 3.33: Fake Sentence, output Bert 80,000 Sentences 10 Epochs Fake, output Bert 210,000 Sentences 10 Epochs Real*

## Conclusion

In this chapter we define our system CBOW Based On Java System implemented by java language programming and explain its classes and the main methods of it. In same time we talk about two version of Detector System Based On Berts (Bert 80,000 Sentences 10 Epochs, output Bert 210,000 Sentences 10 Epochs) model and its implementation and applied three systems on Questionnaire (look on title [2.8.1](#)). Moreover we compared our obtained results with each other and we infer that Detector System Based On Bert that is trained on 80,000 Sentences 10 Epochs better than the other systems.

# Conclusion

There are many tools that we can use to generate fake news and GPT-2 model is one of those tools. In this thesis we knew that the best language model is GPT-2 and Bert model in side of the power of training, sophisticated architectures and that is due to power machine and the huge size of dataset. We knew also that those models are one of the best models to detect fake news.

In this thesis we provide a layer to pre trained model to avoid some problems such as the recurrent of words and uncompleted sentences just to generate a good fake sentences. And in final chapter we used Bert to predict our generated fake news. In the same time we implemented simple system using java programming language and compared it with Arabic Bert to get how strong BERT model is.

# Bibliography

- Aghakhani, H., Machiry, A., Nilizadeh, S., Kruegel, C., and Vigna, G. (2018). Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 89–95. IEEE.
- AKHOLI (2020). gpt2-small-arabic. <https://huggingface.co/akhooli/gpt2-small-arabic>.
- Alammar, J. (2019). The illustrated gpt-2 (visualizing transformer language models). <http://jalammar.github.io/illustrated-gpt2/>.
- Alessia, D., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3).
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ferrero, J., Agnes, F., Besacier, L., and Schwab, D. (2017). Using word embedding for cross-language plagiarism detection. *arXiv preprint arXiv:1702.03082*.
- Kenter, T., Borisov, A., and De Rijke, M. (2016). Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.
- Khorsi, A., Cherroun, H., Schwab, D., et al. (2018). A two-level plagiarism detection system for arabic documents. *Cybernetics and Information Technologies*, 20.
- Lee, J.-S. and Hsiang, J. (2019). Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*.
- Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308.
- Levy, O. and Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

- Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
- Michie, D., Spiegelhalter, D. J., Taylor, C., et al. (1994). Machine learning. *Neural and Statistical Classification*, 13(1994):1–298.
- Müller, A. C., Guido, S., et al. (2016). *Introduction to machine learning with Python: a guide for data scientists*. ” O’Reilly Media, Inc.”.
- Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 2018. Determination press San Francisco, CA.
- Phi, M. (2018). Illustrated guide to lstm’s and gru’s: A step by step explanation.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.
- Richardson, L. (2007). Beautiful soup documentation. *Dosegljivo*: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>. [Dostopano: 7. 7. 2018].
- Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach.
- Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Shi, Z., Chen, X., Qiu, X., and Huang, X. (2018). Toward diverse text generation with inverse reinforcement learning. *arXiv preprint arXiv:1804.11258*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Wang, Q., Xu, J., Chen, H., and He, B. (2017). Two improved continuous bag-of-word models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2851–2856. IEEE.
- Wang, W. Y. (2017). ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Yinfei Yang, C. T. (2018). Advances in semantic textual similarity. <https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html>.
- Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9054–9065.
- Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2019). Dive into deep learning. *Unpublished Draft*. Retrieved, 19:2019.
- Zhou, X. and Zafarani, R. (2018). A survey of fake news: Fundamental theories, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*.