

AMERICAN UNIVERSITY OF BEIRUT

SARABERT:AFFIXING INTER-SENTENCE
TRANSFORMERS TO ARABERT FOR
EXTRACTIVE SUMMARIZATION

by

SAMI BILAL SHAMES EL DEEN

A thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Graduate Program in Computational Science
of Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
May 2022

AMERICAN UNIVERSITY OF BEIRUT

SARABERT:AFFIXING INTER-SENTENCE
TRANSFORMERS TO ARABERT FOR
EXTRACTIVE SUMMARIZATION

by

SAMI BILAL SHAMESELDEEN

Approved by:



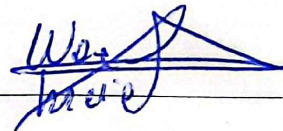
Prof. Mariette Awad, Associate Professor
Electrical and Computer Engineering

Advisor



Dr. Shady Elbassuoni, Associate Professor
Computer Science

Member of Committee



Prof. Wael Khreich, Assistant Professor
Business Information Decision Systems

Member of Committee

Date of thesis defense: May 2, 2022

AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: Shameseldeen Sami Bilal
Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

___ As of the date of submission of my thesis

☒ After 1 year from the date of submission of my thesis .

___ After 2 years from the date of submission of my thesis .

___ After 3 years from the date of submission of my thesis .

Sf.
Signature

12 May 2022
Date

Abstract of the Thesis of

Sami Bilal Shames Ell Deen for Masters of Science
Major: Computational Science

Title: SARA-BERT: Affixing Inter-Sentence Transformers to AraBERT
for Extractive Summarization

Natural language processing (NLP) has made remarkable advancement with the advent of deep learning technology. The deep learning models have produced enhanced results in NLP tasks such as text summarization, text translation, and sentiment analysis. In particular, text summarization is becoming an important task as the number and volume of electronic documents are increasing rapidly. However, NLP for Modern Standard Arabic (MSA) did not witness enough research due to the many challenges the language faces, the complexity of the language itself and the lack of structured data. In this research, we introduce SARA-BERT, an enhanced version of AraBERT that adds inter-sentence transformer layers for extractive summarization tasks. To ensure that the summaries generated achieve a high coverage of the document's main ideas, we propose Semantic Siamese Similarity, a novel evaluation metric that measures the level of similarity between two text inputs. Testing using BLEU, ROUGE, and Semantic Siamese similarity on Sarabert and published related models showed the effectiveness of our proposed model and motivate follow on research.

Table of Contents

ACKNOWLEDGEMENTS	1
ABSTRACT	2
1 Introduction	7
2 Literature Review	9
2.1 Arabic Language	9
2.2 Arabic challenges	10
2.2.1 Morphological richness	10
2.2.2 Orthographic ambiguity	10
2.2.3 Dialectal variation	11
2.2.4 Orthographic inconsistency	11
2.2.5 Resource Poverty	11
2.3 NLP	12
2.3.1 Text Summarization	12
2.4 Previous Work	12
2.4.1 Classical Approaches	13
2.4.2 Machine Learning Approach	13
2.4.3 Other Approaches	14
2.5 Readability Metrics	15
2.5.1 Flesch Reading Ease	15
2.5.2 SMOG Index	15
2.5.3 FOG Index	15
2.5.4 OSMAN	15
2.6 Evaluation Metrics	16
2.6.1 ROUGE	16
2.6.2 BLEU	17
2.6.3 BertScore	17
2.6.4 Triangle Similarity - Sector Similarity (TS-SS)	18
3 Methodology	20
3.1 SaraBERT	20
3.2 Encoder	23

3.2.1 Simple Classifier	23
3.2.2 Recurrent Network	23
3.2.3 Transformer	24
3.3 Siamese Semantic Similarity (SSS)	25
4 Experiments	28
4.1 Dataset	28
4.1.1 CNN/ Daily Mail	28
4.1.2 Dataset Translation	29
4.1.3 Corpus Topic Categories	30
4.1.4 Kalimat Dataset	33
4.2 Data Preprocessing	33
4.3 Training Details	34
4.4 Results	36
4.5 Discussion	38
5 Conclusion	39
A Siamese Semantic Similarity Experiments	40
B Sample Summaries form SaraBert	43
C Translation Quality	45
Bibliography	48

Illustrations

3.1	Overview for the architecture of the original BERT model	21
3.2	Overview for the architecture of the SARaBERT model	22
4.1	Sample English passage translated to Arabic and Diacritized . . .	30
4.2	Histogram of readability comparisons between english and arabic with and without diacritics	31
4.3	Bar chart showing the distribution of document labels per category	32
4.4	Bar chart showing the distribution of corpus sentences and words per category	32
4.5	Cosine Similarity of named entity counts between abstracts and articles	33
4.6	Sample representation of tokenizing and segmenting an arabic passage	34
4.7	Loss training curve of each model	36
4.8	Pearson Correlation between the similarity metrics	38
B.1	Sample SaraBert summarization. Yellow highlighting represents summary extraction using MLP as encoder, Green represents RNN encoder, and Blue represents Transformer	43
B.2	Sample SaraBert+BiLSTM of a good summarization	44
B.3	Sample SaraBert+BiLSTM of a bad summarization	44
C.1	Sample English passage with Fleach score of 11.38 (Hard to read)	45
C.2	Sample Arabic passage (translation of Figure[C.1]) with Osman score of 66.73 (Slight hard to read)	46
C.3	Sample English passage with Fleach score of 83.69 (Easy to read)	46
C.4	Sample Arabic passage (translation of Figure[C.3]) with Osman score of 88.01 (Easy to read)	46

Tables

3.1	interpretation of SSS scores	26
4.1	Document highlighted counts	29
4.2	Average number of tokens for each feature	29
4.3	Average Readability Measures	29
4.4	Translation Experiment Results	29
4.5	Average Readability Measures for Arabic	30
4.6	Distribution of documents between different categories	32
4.7	Distribution of documents between the split sets	35
4.8	Hyper-parameters of different encoders	35
4.9	BLEU scores over Kalimat Dataset	37
4.10	ROUGE scores over Kalimat Dataset	37
4.11	Results of different models under several evaluation metrics . . .	37
A.1	List of a reference sentence (1) and set of candidate sentences(2 to 19)	41
A.2	Results of computing SSS and BertScore (BS) along with other required computations for the candidates in Table[A.1] linked by ID	42
C.2	Correlation between the different features	47
C.1	Average value for documents computed over 1000 sample docu- ments	47

Chapter 1

Introduction

As the amount of textual information available online grows rapidly, it becomes difficult for readers to read large amounts of text and find out which of these texts are useful. As a result, researchers in the field of automatic text summarization could make use of tools that can make multiple-document reading more efficient. The task of text summarization is considered one of the most important and challenging NLP tasks. This task is based on generating short text from long text, so the short text contains the most important information from the original text. The task is often divided into two paradigms known as abstractive summarization and extractive summarization. The first methodology determines the essential sections of the text using statistical tool. The summary is represented by truncating and connecting these sections. The second methodology emulates human activity in summarizing, which is based on presenting the text's core idea in a new linguistic style and using different terms. It incorporates more complicated procedures including paraphrase, generalization, and reordering [1]. This thesis focuses on extractive summarization. While there is a wealth of scientific research in the area of extractive summarization, most of this research is based on English texts, and there is a lack of research on summarizing Arabic texts.

Arabic natural language processing (NLP) is considered more complex than English and other European languages. The main reason for this complexity is the highly derived and rich form of Arabic morphology. That's why Arabic NLP impede research progress compared to other languages. Some examples of these challenges are:

- 1) Morphological richness language [2][3]: Arabic is heavily derived, inflected and has a significant impact on NLP tasks such as stemming and lemming.
- 2) Absence of diacritics: diacritics play an important role in determining the meaning of words and facilitating the task of tokenizing and parsing text.
- 3) No capitalization in Arabic language: without the usage of uppercase letters in Arabic, it will be difficult to identify proper nouns, titles, and abbreviations.

viations.

In this work, we introduce SAraBERT, a novel and enhanced version of AraBERT that adds inter-sentence transformer layers for extractive summarization tasks. To ensure that the summaries generated achieve a high coverage of the document’s main ideas, we also propose a novel evaluation metric, Semantic Siamese Similarity that measures the level of similarity between two text inputs using contextualized embeddings in addition to exact match. We use an English benchmark dataset (CNN/Daily Mail) and translate it to Modern Standard Arabic language to train SAraBERT. Testing the performance of Sarabert in comparison with other published models was evaluated using BLEU, ROUGE, and Semantic Siamese similarity on Kalimat Dataset. Results have shown the effectiveness of our proposed model.

In summary, the main contributions of this thesis, which are as follows:

- A novel Arabic language model for extractive summarization that builds on AraBERT model.
- A novel hybrid similarity metric that measures similarities between 2 text based on the semantic and syntactic features.
- An arabic corpus translated from English that targets the extractive summarization and Question Answering tasks, available upon request.

In the next chapter [2] we survey the related work and background information required for this work. Chapter[3] clearly states the model’s architecture and explains the proposed metric. Chapter [4] shows the experimental research and highlights the obtained results along with some insights. Finally Chapter [5] concludes the thesis with follow on research directions.

Chapter 2

Literature Review

The chapter summarizes the related literature. It introduces the Arabic language and lists its challenges. The chapter then mentions previous work done on text summarization for Arabic documents. It also explains the metrics used for evaluating the readability level of the input documents and the ones used for evaluating the generated summaries.

2.1 Arabic Language

Arabic is one of the most spoken languages in the world. Arabic relates to Islam and more than 200 million Muslims pray five times a day utilizing this dialect. Furthermore, Arabic is the official first language of many Arab countries and is among the official languages of the United Nations. Arabic is an extremely rich language and is related to the Semitic dialects, which is different from the Indo-European languages spoken in the West.

Arabic language has a rich complex grammatical structure[4]. For instance, a noun and its modifiers need to agree in number, gender, case, and definiteness [5]. Moreover, in Arabic, there are advancements that really mean “Mother of” or “Father of” to show ownership, a trademark, or a property, and use gendered pronouns; it has no fair-minded pronouns[6]. Arabic sentences can be nominal (subject-verb), or verbal (verb-subject) with free order; however, English sentences are fundamentally in the (subject-verb) order. The free order property of the Arabic language presents a crucial challenge for some Arabic NLP applications[7] (Check section 2.2).

Three types characterize Arabic: Classical Arabic (Quranic), Modern Standard Arabic (MSA), and Dialectal Arabic[8][9]. Classical Arabic is mainly spoken in Arab-speaking countries, it is found in religious writings such as Sunnah and Hadith, and in many historical documents[10]. Diacritic marks are commonly used within Classical Arabic as phonetic guides to show the correct pronunciation. In contrast, diacritics are considered optional in most other Arabic scripts [11]. MSA is used for television, newspapers, poetry, and

books. Arabic lessons at the Arab Academy are also taught in a modern standard format. MSA can be transformed to accommodate new words that must be created due to science or technology. However, Arabic writing has not changed its alphabet, spelling, or vocabulary for at least four millennia.

In this thesis, illustrative examples are used for clarification. Examples are given in MSA as it represents the majority of written material and formal documents, lectures, and articles. Moreover, it is the universal language version that all Arabic speakers can understand.

2.2 Arabic challenges

Arabic grammar has a rich morphology and intricate sentence structure and grammarians have described it as the language of dad ("لغة الضاد") [12] [13] states that Arabic has a greatly rich morphology mixing templatic and affixational morphemes as well as complex morphological norms.

In what follows, we list as highlights, the number of modeling challenges for arabic NLP.

2.2.1 Morphological richness

Arabic words have numerous forms resulting from a rich inflectional system that includes features for gender, number, person, aspect, mood, case, and several attachable clitics. As a result, it is not uncommon to find single Arabic words that translate into five-word English sentences (وَسَيَذُرُونَهَا) "wa-sa-ya-drus-uuna-ha' "and they will study it'. This challenge leads to a higher number of unique vocabulary types compared to English, which is challenging for machine learning models.

2.2.2 Orthographic ambiguity

The Arabic script uses optional diacritical marks to represent short vowels and other phonological information that is important to distinguish words from each other. These marks are almost never used outside of religious texts and children's literature, which leads to a high degree of ambiguity. Arab scholars do not usually have a problem with reading undiacritized Arabic, but it is a challenge for Arabic learners and computers. This out-of-context ambiguity in Standard Arabic could lead on average to a staggering 12 analyses per word:

for example, the readings of the word كَتَبْتُ "ktbt" (no diacritics) includes:

- كَتَبْتُ - katabtu – I wrote
- كَتَبَتْ - katabat – she wrote
- كَتَبْتُ - ka+t ibit – such as Tiber

2.2.3 Dialectal variation

Arabic is also not a single language but rather a family of historically linked varieties, among which MSA is the official language of governance, education, and the media, while the other varieties, so-called dialects, are the languages of daily use in spoken, and increasingly written, form. Arab children grow up learning their native dialects, such as Egyptian, Levantine, Gulf, or Moroccan Arabic, which have their own grammars and lexicons that differ from each other and from MSA. For example, the word for ‘car’ is (سيارة) sayyaara in MSA, (عربية) arabiyya in Egyptian Arabic, (كرهبة) karhba in Tunisian Arabic, and (موتَر) motar in Gulf Arabic. The difference can be significant to a point that using MSA tools on dialectal Arabic leads to quite sub-optimal performance.

2.2.4 Orthographic inconsistency

MSA and dialectal Arabic are both written with high degree of spelling inconsistency, especially on social media: A third of all words in MSA comments online having spelling errors; and dialectal Arabic has no official spelling standards, although there are efforts to develop such standards computationally, such as the work on Conventional Orthography for Dialectal Arabic (CODA). Furthermore, Arabic can be encountered online written in other scripts, most notably, a romanized version called Arabizi that attempts to capture the phonology of the words.

2.2.5 Resource Poverty

Data is the bottleneck of Arabic NLP; this is true for rule-based approaches that need lexicons and carefully created rules, and for machine learning ap-

proaches that need corpora and annotated corpora, Although Arabic un-annotated text corpora are quite plentiful, Arabic morphological analyzers and lexicons as well as annotated and parallel data in non-news genre and in dialects are limited. While none of the issues mentioned here are unique to Arabic; as Turkish and Finnish are morphologically rich and Hebrew is orthographically have dialectal variants for instance, the combination and degree of these phenomena in Arabic creates a particularly challenging situation for NLP research and development. Additional information has been published on Arabic computational processing challenges[14][15].

2.3 NLP

NLP is a branch of computer science that attempts to make it easier for machines (computers that comprehend machine language or programming languages) and humans to interact (who communicate and understand natural languages like English, Arabic and Chinese etc.) using written languages. Today, many applications we use and rely on, are powered by NLP.

2.3.1 *Text Summarization*

With all of its recent advancements, the task of text summarization is one of the most critical problems that computer capabilities confront. The goal of this task is to create short text from larger text that contains the most relevant information from the original text. There are two basic methodologies used to summarize the texts which are extractive summarization from which most systems with good results came out, and abstractive summarization that simulate human summarization. The first methodology is based on using a statistical approach to determine the essential sections of the text, similar to Brkibir et al's work [16]. In a semantic approach, such as Imam's work on the Arabic language[17], the summary is represented by truncating and connecting these sections, similar to Knight et al's work on the English language [18]. The second methodology is focused on mimicking human activity in summarizing, which is based on presenting the text's core idea in a new linguistic style and using different terms; it incorporates more complicated procedures including paraphrase, generalization, and reordering [1]. Previous research has attempted to generate abstract summaries utilizing linguistically inspired restrictions [19] or by transforming the incoming material syntactically [20][21].

2.4 Previous Work

In order to provide the necessary context to the proposed solution, it is worth investigating previous research, and identifying the pros and cons of each

approach.

2.4.1 Classical Approaches

Statistical methods are widely used in text summarization which are based on the concept of relevance score and Bayesian classifiers [22]. Word Frequency approach is the most used methodology for sentence scoring where the sentence score is calculated by the sum of the frequencies while avoiding all stop words. The proposed solution in [23] generate news titles by combining Word-Frequency, sentence position and similarity measures methodologies. Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical methodology that represents the importance of a word to document in a collection of documents (corps) [24]. TF-IDF is an improvement on the Word Frequency and shows how the weights are distributed on the words of a document. TF-IDF is used frequently in auto-summarization systems [25] [26] [27] [28]. An Arabic summarization system called AATSS [29] adopted an extractive text summarization approach that was mainly based on sentence weighting and scoring. However, it highly depends on the size of the document in terms of number of sentences, which leads to generation of improper grammar context, and it lacks automatic Arabic entity recognition system and Arabic pronoun resolution system. The classical statistical approaches are used for both single and multi-document summarization and can be used to enhance the selection of important sentences or the elimination of redundant sentences but it fails to understand the text, since it only depends on statistical measures [30]. [31] proposed a solution for English language by tokenizing the text into clean sentences, then pass the sentences into BERT model to generate embeddings. Then using K-means they generated clusters and selected summary sentences based on closest sentence embedding to each cluster centroid.

2.4.2 Machine Learning Approach

Text summarization is considered as a binary classification problem, where a set of documents and their extractive summaries are used as a training set, and each sentence is classified as a summary sentence or non-summary based on statistical, semantic features or a combination of them [32] [16] [33] [34]. Machine learning methods can be suitable for document summarization as several methods have been introduced and proven to be effective in performing summarization tasks. We will discuss sequence to sequence models and encoder decoder architectures in what follows.

Sequence-to-Sequence Model

The neural abstractive summarization with sequence-to-sequence models emerged in [35] [36]. This approach has been applied to tasks such as headline generation [37] and article summarization [38]. Chopra et al. [39] show that

attention approaches that are more specific to summarization can further improve the performance of models. Molham et al. [40] re-implemented the sequence-to-sequence framework on the Arabic language, which has not witnessed the employment of this model in the text summarization before. However, the authors state that the work still requires expanding the dataset to cover more articles, and to infer new models that are beneficial with the Arabic language since it has a unique grammatical language written from right to left.

Encoder-Decoder Model

Google AI pre-trained language model called Bidirectional Encoder Representations from Transformers (BERT), proved highly efficient at language understanding and achieved convincing results in most NLP tasks. The Arabic language model ARABERT based on BERT for Arabic language was evaluated on Named Entity Recognition, Sentiment Analysis and Question Answering [41]. Abdulla et al. [42] proposed an extractive Arabic text summarizer based on ARABERT to summarize the Arabic documents by evaluating and extracting the most important sentences at a document. This proposed approach generated a good summary by extracting the most important sentences from paragraphs and generated a coherent and meaningful summary. However, the model highly depends on sentence boundaries, its coverage accuracy decreases when the text is too long, and extracted sentences sometimes contain linguistic expressions that creates ambiguous summaries.

2.4.3 Other Approaches

Researchers et al. of [43] managed to solve the problem of sentence boundary detection by adding an additional layer in the embedding that specifies the beginning and end of each consecutive sentence inserted into the encoder.

Go et al. [44] explored the effects of language variants, data sizes, and fine-tuning task types in Arabic pre-trained language models. The results suggested that the variant proximity of pre-training data to fine-tuning data is more important than the pre-training data size. Moreover, other design decisions can be explored that may contribute to the fine-tuning performance, including vocabulary size, tokenization techniques, and additional data mixtures.

Nasser et al. [45] presented an LSTM-based morphological disambiguation system for Arabic which has significantly outperformed the state-of-the-art systems. The paper suggested exploring additional deep learning architectures for morphological modeling and disambiguation, especially sequence-to-sequence models. It also suggested to further investigate the role of syntax features in morphological disambiguation, and explore additional techniques for more accurate tagging.

2.5 Readability Metrics

In order to understand the level of complexity in reading and understanding a document from the corpus, readability metrics were used to measure the comprehension of written text by readers of different levels of education. Readability focuses on multiple factors such as length of sentences and words, number of difficult and complex words, and the number of syllables in each word. A lot of research on such metrics for the English language have emerged, we will mention and formalize a few of them:

2.5.1 Flesch Reading Ease

Flesch Ease [46] is a readability test designed to indicate how difficult a passage in English is to understand ranging from 0 indicating the passage to be very confusing to 100 indicating it is very easy to read.

$$Flesch = 206.8351.015 \left(\frac{total\ words}{total\ sentences} \right) - 84.6 \left(\frac{total\ syllables}{total\ words} \right) \quad (2.1)$$

2.5.2 SMOG Index

SMOG Index [47] is used to measure how many years of education the average person needs to have to understand a text.

$$SMOG = 1.043 \sqrt{\frac{total\ polysyllables \times 30}{total\ sentences}} + 3.1291 \quad (2.2)$$

2.5.3 FOG Index

The FOG index [48] estimates the years of formal education a person needs to understand the text on the first reading.

$$Fog = 0.4 \left(\frac{total\ words}{total\ sentences} + \frac{complex\ words}{total\ words} \right) \quad (2.3)$$

2.5.4 OSMAN

OSMAN [49] is a modified version of the conventional readability formulas such as Flesch and Fog. Since Arabic is highly inflectional and derivational, word length and number of sentences cannot be sufficient indicators of text readability. Additional factors should be considered such as "Faseeh words" which are words containing

(ء، ئ، وُ، ذ، ظ، ون، وا، ح)

$$OSMAN = 200.791 - 1.015 \times \left(\frac{total\ words}{total\ sentences} \right) - 24.181 \times \frac{total\ long\ words + syllables\ in\ word + complex\ words + Faseeh\ words}{total\ words} \quad (2.4)$$

2.6 Evaluation Metrics

2.6.1 ROUGE

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [50] is widely used for performance evaluation of summarization techniques. Some of the ROUGE methods are commonly used to measure performance of the model in summarizing documents. $ROUGE_N$ is the N-gram ($N \geq 1$) recollection between a system summary and human-generated or reference summaries. It is used to estimate the fluency of the summaries.

$$Rouge_N = \frac{\sum_{s \in refsum} \sum_{gram_n \in S} count_{match}(gram_n)}{\sum_{s \in refsum} \sum_{gram_n \in S} count(gram_n)} \quad (2.5)$$

$ROUGE_L$ is used to identify the longest co-occurring in sequence N-grams automatically. Assume A is the set of sentences of the reference summary and B is the set of sentences of the candidate summary represented by the sequence of words and LCS-based F score (F_{lcs}) indicates the similarity between A (of length m) and B (of length n):

$$R_{lcs} = \frac{LCS(A, B)}{n} \quad (2.6)$$

$$P_{lcs} = \frac{LCS(A, B)}{m} \quad (2.7)$$

$$F_{lcs} = \frac{2R_{lcs} \times P_{lcs}}{R_{lcs} + P_{lcs}} \quad (2.8)$$

where $LCS(A, B)$ denotes the length is the LCS of A and B

2.6.2 BLEU

BiLingual Evaluation Understudy (BLEU) is a lexical-based metric that is mostly used for automatic evaluation of machine-translated text. The BLEU score is a number between 0 and 1 that measures the similarity between machine translated text and a set of high quality reference translations. A value of 0 means that the machine translation does not overlap the reference translation (low quality), and a value of 1 means that the reference translation matches exactly (high quality).

$$BLEU = \min \left(1, \exp \left(1 - \frac{\text{reference} - \text{length}}{\text{output} - \text{length}} \right) \right) \left(\prod_{i=1}^4 \text{precision}_i \right)^{1/4} \quad (2.9)$$

with

$$\text{precision}_i = \frac{\sum_{\text{snt} \in \text{Cand-Corpus}} \sum_{i \in \text{snt}} \min(m_{\text{cand}}^i, m_{\text{ref}}^i)}{w_t^i = \sum_{\text{snt}' \in \text{Cand-Corpus}} \sum_{i' \in \text{snt}'} m_{\text{cand}}^{i'}} \quad (2.10)$$

where

1. m_{cand}^i is the count of i-gram in candidate matching the reference translation
2. m_{ref}^i is the count of i-gram in the reference translation
3. w_t^i is the total number of i-grams in candidate translation

2.6.3 BertScore

BERTScore [51] leverages the pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences using cosine similarity between the tokens' embeddings. BERTScore addresses a common problem in n-gram-based metrics. the fail to robustly match phrases as those metrics can only cover the grammar of the sentence without interpreting the semantic meaning, this leads to performance underestimation when semantically-close phrases are penalized because they differ on the word matching level with the reference.

Let $x = \langle x_1, \dots, x_k \rangle$ be the set of embedding vectors the model generates on a tokenized reference sentence. Similarly for the candidate sentence $\hat{x} = \langle \hat{x}_1, \dots, \hat{x}_k \rangle$. BERTScore can be applied for obtaining Recall Precision, and F1 score:

$$R_{BERT} = \frac{\sum_{x_i \in x} \text{idf}(x_i) \max_{\hat{x}} x_i^\top \hat{x}_j}{\sum_{x_i \in x} \text{idf}(x_i)} \quad (2.11)$$

$$P_{BERT} = \frac{\sum_{x_i \in x} \text{idf}(\hat{x}_i) \max_{\hat{x}} \hat{x}_i^\top x_j}{\sum_{x_i \in x} \text{idf}(\hat{x}_i)} \quad (2.12)$$

$$F_{BERT} = 2 \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (2.13)$$

2.6.4 Triangle Similarity - Sector Similarity (TS-SS)

TS-SS [52] is a geometric approach to measure the similarity level among documents. It considers documents to be represented as vectors and measures similarity based on the angle and Euclidean distance between each pair. We briefly explain some of the popular geometric similarity measures that we will be using.

Cosine Similarity: Cosine similarity computes pairwise similarity between 2 documents using dot product and magnitude of document vectors A and B in high-dimensional space [53]

$$\text{cosine}(A, B) = \frac{\sum_{n=1}^k A(n) \cdot B(n)}{|A| \cdot |B|} \quad (2.14)$$

The resulting similarity ranges from 0 to 1, with 1 meaning the 2 vectors are overlapped and the 2 documents have maximum similarity ($\cos(0) = 1$), and 0 meaning that the documents have no similarity commons.

Euclidean Distance: ED is another geometrical measure used to measure similarity of 2 documents. Each document is represented as a point in space based on the number of features in a vector representation. ED computes difference between 2 points in n dimensional space based on their coordinate using the following equation:

$$ED(A, B) = \sqrt{\sum_{n=1}^k (A(n) - B(n))^2} \quad (2.15)$$

Using ED, highest similarity between 2 vectors happens when they are plotted in the same position in space, making the distance difference between them 0.

Triangle's Area Similarity (TS): The Triangle Area Similarity combines 3 geometric similarity characteristics, the angle between vectors, magnitude of vectors, and ED.

$$TS(A, B) = \frac{|A| \cdot |B| \cdot \sin(\theta')}{2} \quad (2.16)$$

where $\theta' = \cos^{-1}(V) + 10$ $V = \text{cosineSim}(A, B)$

If the vectors A, B are overlapping, no area can be computed thus to overcome this problem we increase θ' by 10

Sector's Area Similarity (SS): Merely TS alone is not robust enough to interpolate vectors' differentiation precisely to produce accurate similarity results due to missing components. Thus the SS has been introduced:

$$SS(A, B) = \pi \cdot (ED(A, B) + ||A| - |B||)^2 \cdot \frac{\theta'}{360} \quad (2.17)$$

The final formula of the TS-SS metric is produced by combining TS and SS

by multiplying them together.

$$TS - SS(A, B) = TS(A, B) \times SS(A, B) \quad (2.18)$$

Chapter 3

Methodology

This chapter presents the model architecture built for the translation task and its different encoder extensions. The chapter ends with introducing our new hybrid metric (SSS) and showcases its implementation and algorithm.

3.1 SAraBERT

Let d denote a document containing several sentences $[sent_1, sent_2, \dots, sent_m]$, where $sent_i$ is the i^{th} sentence in the document. Extractive summarization can be defined as the task of assigning a label $y_i \in \{0, 1\}$ to each $sent_i$, indicating whether the sentence should be included in the summary or not. Thus, if $y_i = 1$ then the sentence is considered in the summary due to its importance in terms of document content. BERT as shown in Figure[3.1] is considered the model of choice because it showed better performance compared to other NLP statement embedding algorithms. To keep the original BERT pre-training objective, AraBERT uses Masked Language Modeling (MLM) to improve the pre-training process by letting the model predict the entire word rather than receiving clues from part of the word. In addition, it also uses Next Sentence Prediction (NSP) to help the model understand the relationships between sentences [41]. To achieve sentence ranking in the document based on content importance, we will apply modifications to the embedding layers as shown in Figure[3.2] to highlight sentence endpoints for the model to rank each sentence independently.

Encoding Multiple Sentences: We insert $[CLS]$ token before each sentence and a $[SEP]$ token after each sentence. The $[CLS]$ is used as a symbol to represent the start of a sequence and the $[SEP]$ indicated the end of that sequence. Inserting multiple $[CLS]$ tokens help in specifying the boundaries of each sentence.

Interval Segment Embedding: We add another layer of embedding to distinguish multiple sentences within a document. So, for $sent_i$ we assign a segment embedding E_A or E_B based on whether it's even or odd. For ex-

ample, For $[sent_1, sent_2, sent_3, sent_4, sent_5]$ the interval segment layer assign $[E_A, E_B, E_A, E_B, E_A]$. The vector T_i that corresponds to the i^{th} $[CLS]$ token, will be used as the representation for $sent_i$.

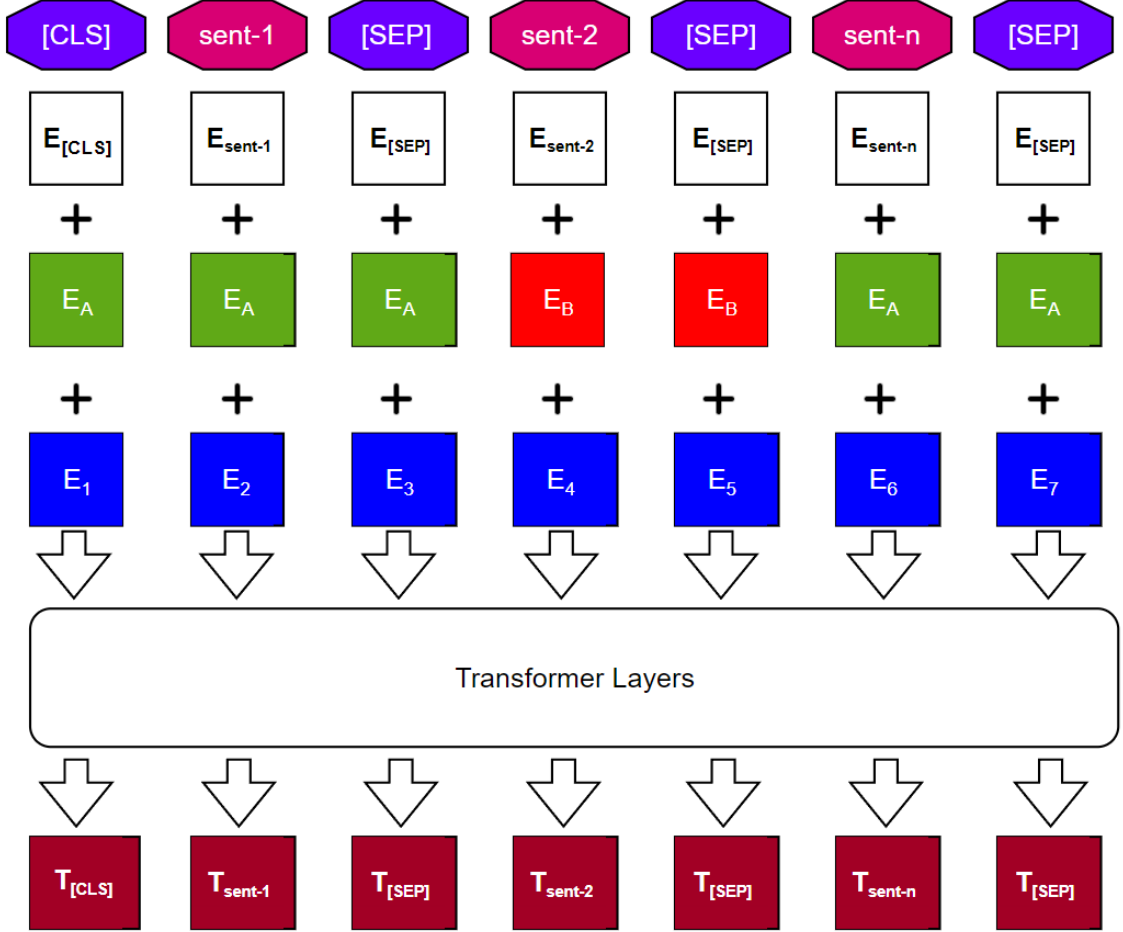


Figure 3.1: Overview for the architecture of the original BERT model

The sequence on top is the input document, followed by the summation of three kinds of embeddings for each token. The summed vectors are used as input embeddings to the transformer layers, generating contextual vectors for each token T_i .

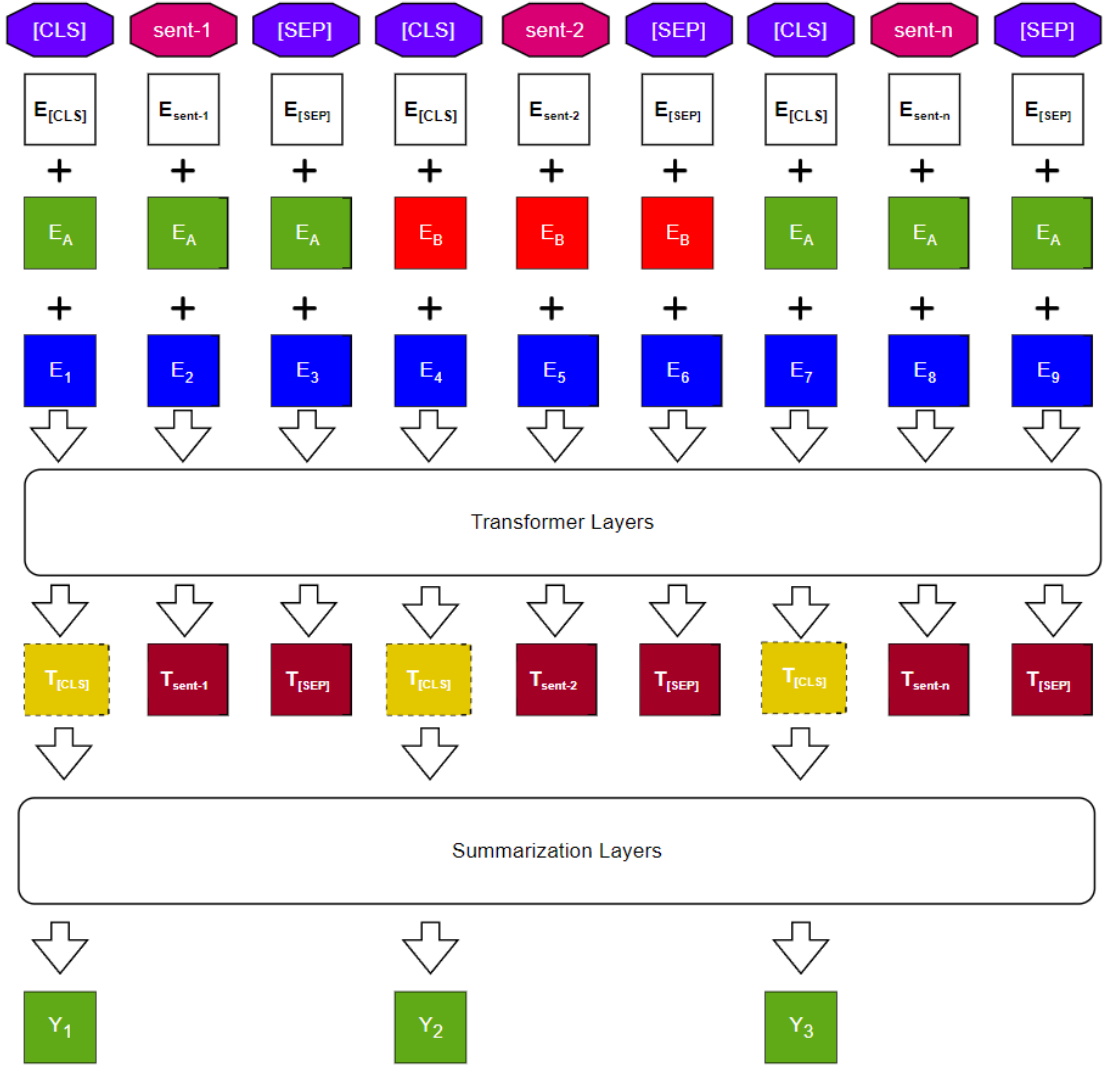


Figure 3.2: Overview for the architecture of the SaraBERT model

The difference between SaraBERT and the model in Figure[3.1] is the insertion of $[CLS]$ tokens before each sentence, and alternating the sentence segmenting embedding values along with the addition of a summarization layer that derives a sentence score from the $[CLS]$ token embeddings.

Algorithm 1 Summarization Model

Input: Text

Output : Sentence Scores

```
tokens ← tokenize(Text)
sentences ← sentence_segmentation(tokens)
tokenID ← array [|sentences|] [maxsent ∈ sentences(|sent|)]
segmentID ← array [|sentences|] [maxsent ∈ sentences(|sent|)]
for i = 1 to |sentences| do
  sentences[i] ← {"[CLS]", sentences[i], "[SEP]"}
  for j = 1 to |sentences[i]| do
    tokenID[i][j] ← getTokenID(sentences[i][j])
    if i is Even then
      segmentID[i][j] ← 0
    else
      segmentID[i][j] ← 1
    end if
  end for
end for
bertEmbedding ← AraBERT(tokenID, segmentID)
sentence_embeddings ← get_CLS_embeddings(bertEmbedding)
sentence_scores ← encode(sentence_embeddings)
```

Next section will discuss the different types of layers that were added to AraBERT to improve the scoring task.

3.2 Encoder

3.2.1 Simple Classifier

We add a linear layer on the AraBERT outputs T_i with sigmoid as the activation function to get the predicted score:

$$\hat{Y}_i = \sigma(W_o T_i + b_o) \quad (3.1)$$

Where σ is a Sigmoid function which is a non-linear activation function used mostly in output layers for predicting probability based outputs (specifically used for binary classification). The Sigmoid function has the following expression:

$$\sigma(x) = \left(\frac{1}{1 + e^{-x}} \right) \quad (3.2)$$

3.2.2 Recurrent Network

We apply an LSTM layer over the BERT outputs to learn summarization-specific features. To stabilize the training, pergate layer normalization Ba et al. [54]

is applied within each LSTM cell. At time step i , the input to the LSTM layer is the AraBERT output T_i , and the output is calculated as follows:

$$\begin{pmatrix} F_i \\ I_i \\ O_i \\ G_i \end{pmatrix} = LN_h(W_h h_{i-1}) + LN_x(W_x T_i) \quad (3.3)$$

$$C_i = \sigma(F_i) \odot C_{i-1} \quad (3.4)$$

$$h_i = \sigma(O_i) \odot \tanh(LN_c(C_i)) \quad (3.5)$$

Where F_i, I_i, O_i are forget gates, input gates, output gates; G_i is the hidden vector and C_i is the memory vector; h_i is the output vector; LN_h, LN_x, LN_c are there difference layer normalization operations; Bias terms are not shown. The final output layer is also a sigmoid classifier:

$$\hat{Y}_i = \sigma(W_o h_i^L + b_o) \quad (3.6)$$

3.2.3 Transformer

Adds more Transformer layers only on sentence representations T_i , extracting document level-features focusing on summarization tasks.

$$\tilde{h}^l = LN(h^{l-1} + MHAtt(h^{l-1})) \quad (3.7)$$

$$h^l = LN(\tilde{h}^l + FFN(\tilde{h}^l)) \quad (3.8)$$

$$h^0 = PosEmb(T) \quad (3.9)$$

$LN(x)$ is a normalization operation [54] that re-centers and re-scales the input x where $x = (x_1, x_2, \dots, x_H)$ is a vector representation of an input of size H the normalization operation being represented as follows:

$$LN(x) = \frac{x - \mu}{\varphi}, \quad \mu = \frac{1}{H} \sum_{i=1}^H x_i, \quad \varphi = \sqrt{\frac{1}{H} \sum_{i=1}^H (x_i - \mu)^2} \quad (3.10)$$

$MHAtt$ is a Multi-Head Attention operation [55]. Before going into how Multi-Head attention work, we should define what is an attention function. An attention function is a mapping from a query matrix Q and a set of key-value pairs (K, V) to an output vector that has a sum of elements equal to one which will represent the percentage of attention each element should have.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad softmax(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^H e^{z_j}} \quad (3.11)$$

The Multi-Head attention allows the model to spread attention over differ-

ent positions.

$$MHAtt(h^l) = Concat(head_1^l, \dots, head_H^l)W^O, \quad head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3.12)$$

Where $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ are trainable parameter matrices.

$FFN(x)$ is a fully connected feed-forward network having two linear transformations with a ReLU activation in between.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (3.13)$$

$PosEmb(T)$ is the function that adds positional embedding to the sentence vectors T . Several choices can be made for the positional embedding, for this work, we will use sine and cosine functions of different frequencies.

$$PoseEmb(T_i) = \begin{cases} \sin(\frac{T_i}{10000^{2i/d_{model}}}) & \text{if } i = 2k \\ \cos(\frac{T_i}{10000^{2i/d_{model}}}) & \text{if } i = 2k + 1 \end{cases} \quad (3.14)$$

The superscript l indicates the depth of the stacked layer with $l = 0$ representing first layer and $l = L$ representing the last layer. The final output layer will still be a sigmoid classifier:

$$\hat{Y}_i = \sigma(W_o h_i^L + b_o) \quad (3.15)$$

3.3 Siamese Semantic Similarity (SSS)

Rouge [50] has been used as a metric for determining the quality of a summary by comparing a candidate summary (generated by machine) to a reference summary (generated by humans). The way it does this projects only the level of similarity on the syntactic level without the coverage of the context. Summaries may cover a large portion of the original context but with fewer words (and more verbose) which Rouge can't keep track of. Language models have the ability to map the context of a passage into a fixed size vector. We can compute the cosine similarity of the 2 embeddings to obtain a similarity measure at the semantic level.

Reference	علي يأكل الطعام في الليل	ROUGE	Cosine Similarity
Candidate	هو يتناول الطعام مساء	0.0	0.9074

Since cosine similarity treats all dimensions equally in the semantic space, and 2 vector points might overlap yet stay very distant from each other, the distance should be added into the evaluation, since the smaller the value is the higher the score should be, we insert the distance difference in the denominator. Moreover, we will wrap it with a square root to slow the growth

speed of the value as the difference increase (this will later become helpful in interpreting the computed values).

$$\frac{\cos_sim(cand_{embedding}, ref_{embedding})}{\sqrt{\|cand_{embedding} - ref_{embedding}\|_2 + 1}} \quad (3.16)$$

A value of 1 was added because the distance difference can be 0, thus to eliminate the possibility of a division by zero error.

Reference	تناول الولد تفاحة	Cos_Sim	L2 norm
Candidate ₁	أكل الطفل قطعة	0.873	5.989
Candidate ₂	أكل الطفل فاكهة	0.867	5.686

Moreover, rouge is still an important component of the formula to keep an eye on the grammar.

$$\frac{\cos_sim(cand_{embedding}, ref_{embedding}) \cdot rouge(cand, ref)}{\sqrt{\|cand_{embedding} - ref_{embedding}\|_2 + 1}} \quad (3.17)$$

It remains the element of frequency to be considered, Two sentences can be similar in context yet one being redundant more than the other.

تناول الولد تفاحة	Cos_Sim	Rouge	L2 norm	FreqDiff
أكل الطفل تفاحة	0.957	0.222	3.637	1.000
الولد تفاحة تفاحة تناول تناول	0.930	0.889	5.005	1.667

Thus a penalty should be added on the difference in frequency distribution among words where we divide the the number of unique words over the number of total words. This step was added because an embedding does not take redundancy and coverage into consideration which can affect level of similarity for tasks as summarizing. The final Formula can be seen in Algorithm[2]

SSS score	Interpretation
0-0.1	no clear resemblance or no similarity at all
0.1-0.2	the gist is clear, but not all context is covered
0.2-0.5	context is significantly covered using different words
≥ 0.5	exact replica with slight modifications

Table 3.1: interpretation of SSS scores

Algorithm 2 Semantic Textual Similarity

Input: Candidate, Reference

Output : SimilarityScore

$rouge_1 \leftarrow rouge_1(\text{Candidate}, \text{Reference})$
 $rouge_2 \leftarrow rouge_2(\text{Candidate}, \text{Reference})$
 $rouge_L \leftarrow rouge_L(\text{Candidate}, \text{Reference})$
 $rouge \leftarrow \frac{rouge_1 + rouge_2 + rouge_L}{3}$
 $embedding_1 \leftarrow AraBERT.encode(\text{Candidate})$
 $embedding_2 \leftarrow AraBERT.encode(\text{Reference})$
 $cosine_similarity \leftarrow \frac{embedding_1 \cdot embedding_2}{\|embedding_1\|_2 \times \|embedding_2\|_2}$
 $DistDiff \leftarrow \|embedding_1 - embedding_2\|_2$
 $FreqDiff_c \leftarrow \frac{unique_words(Candidate)}{total_words(Candidate)}$
 $FreqDiff_r \leftarrow \frac{unique_words(Reference)}{total_words(Reference)}$
 $FreqDiff \leftarrow \frac{\min(FreqDiff_r, FreqDiff_c)}{\max(FreqDiff_r, FreqDiff_c)}$
 $SimilarityScore \leftarrow \frac{cosine_similarity \times rouge \times FreqDiff}{\sqrt{DistDiff + 1}}$

Chapter 4

Experiments

This chapter covers the different experiments that have been designed to investigate the translated dataset quality and performance of SaraBert. Results are discussed and conclusions drawn.

4.1 Dataset

4.1.1 CNN/ *Daily Mail*

We used the CNN / DailyMail dataset to train SaraBERT, it is an English dataset containing over 300,000 unique news articles written by CNN and DailyMail journalists. The dataset can be used to train a model for the task of extractive and abstractive summarization. Each document instance from the dataset consists of 3 components:

1. **id**: A string containing the hexadecimal SHA1 hash of the URL from which the story was obtained.
2. **Article**: A string containing the body of a news article.
3. **Highlights**: A string containing the highlights of the article written by the author of the article.

We computed multiple insights to understand and analyse the dataset on different granular levels. Table[4.1] displays the overall number of documents, sentences, and words the dataset is composed of. Table[4.2] shows the average number of tokens that are found in each article and highlight. Moreover, the distribution of the readability measure is important to look after as it defines whether the passages fed into the model while training are naive or not. Table[4.5] illustrates the average readability level of the documents in the English dataset. The results show that the passages are not for beginner-level readers and requires educational background of at least 3 years to fully understand the context of the document.

Dataset Analytics	Count
Number of Documents	211,935
Number of Sentences	17,253,935
Number of Words	140,342,917

Table 4.1: Document highlighted counts

Feature	Mean Token Count
Article	781
Highlights	56

Table 4.2: Average number of tokens for each feature

Metric	Value (Mean \pm std)
Flesch Reading Ease	49.1 \pm 13.02
SMOG Index	3.44 \pm 5.7
Gunning FOG Index	13.42 \pm 3.58

Table 4.3: Average Readability Measures

4.1.2 Dataset Translation

Since our model should be trained on an Arabic dataset, we examined state-of-the-art machine translation model mbart [56] along with Google-Trans¹ service. We evaluated each of the models on the Opus dataset [57]. Results are shown in Table[4.4]. We can see competitive results from both translators, so the tie breaker was the duration required to translate a document.

	BLEU	METEOR	Translation Duration (Docs/minute)
GoogleTrans	0.413 \pm 0.386	0.24	30
mBart	0.413 \pm 0.382	0.25	20

Table 4.4: Translation Experiment Results

We also generated diacriticisation for the Arabic documents [58] to see whether the diacritics have an effect on the readability level or not. Table[4.5] and Figure[4.7] shows that the documents containing diacritics have closer readability estimations to English than documents without diacritics but it will not be a problem since [49] explained that this behavior is normal and what is important is that the correlation between Flesch and OSMAN (without diacritics) is high.

We investigated if translation will affect the readability of a document and whether large difference between the English and Arabic readabilities of the

¹pypi.org/googletrans

English	Arabic	Arabic with Diacritics
<p>Timothy Bradley says he needs to beat Manny Pacquiao for a second time in Las Vegas on Saturday to move on from the controversial conclusion of their first fight two years ago. The WBO welterweight champion won a contentious points decision when the pair met in June 2012, inflicting a first defeat on Pacquiao in seven years. Boxing commentators roundly criticized the result while former heavyweight champion Lennox Lewis said the scoring showed that boxing had lost its integrity.</p>	<p>يقول تيموثي برادلي إنه يحتاج إلى التغلب على ماني باكويو للمرة الثانية في لاس فيجاس يوم السبت للمضي قدما من الاستنتاج المثير للجدل لأول معركة لها قبل عامين. فاز بطل WBO Welterweight بقرار نقاط مثيرة للجدل عندما اجتمع الزوج في يونيو 2012، مما ألحق بالهزيمة الأولى على باكويو في سبع سنوات. انتقد المعلقون الملاكمة بشكل مستدير النتيجة بينما قال لينوكس لويس بطل الوزن الثقيل السابق إن التهديد أظهر أن الملاكمة فقدت سلامتها.</p>	<p>يقول تيموثي برادلي إنه يحتاج إلى التغلب على ماني باكويو للمرة الثانية في لاس فيجاس يوم السبت للمضي قدما من الاستنتاج المثير للجدل لأول معركة لها قبل عامين. فاز بطل WBO Welterweight بقرار نقاط مثيرة للجدل عندما اجتمع الزوج في يونيو 2012، مما ألحق بالهزيمة الأولى على باكويو في سبع سنوات. انتقد المعلقون الملاكمة بشكل مستدير النتيجة بينما قال لينوكس لويس بطل الوزن الثقيل السابق إن التهديد أظهر أن الملاكمة فقدت سلامتها.</p>

Figure 4.1: Sample English passage translated to Arabic and Diacritized

original and translated documents will influence our model. After experimenting on 1000 translated samples, and computing the ROUGE score of the resulting summaries along with the readability of the original and translated samples, results have shown that the level of readability does not affect the model’s output and there is no correlation between readability difference and ROUGE. More details are provided in Appendix[C].

Metric	Value (Mean \pm std)
OSMAN - Diacritics	78.45 \pm 5.18
OSMAN + Diacritics	70.28 \pm 8.95

Table 4.5: Average Readability Measures for Arabic

4.1.3 Corpus Topic Categories

We categorised the content of the corpus using a text classification model originated from [59]. The model $f(w, c)$ returns a value v between 0 and 1 representing the degree of confidence that the word w belongs to category c . For example, $f(\text{apple}, \text{food}) = 0.93$ is read as "The model f is 93% confident that apple belongs to category food". Let $\vec{v} = g(w)$ where \vec{v} is the confidence vector of w having category c_i ’s confidence value assigned to a fixed position i in the output vector \vec{v} . In this experiment, the 8 positions respectively correspond to:

1. Art and Photography
2. Beauty and Fashion
3. Business and Finance

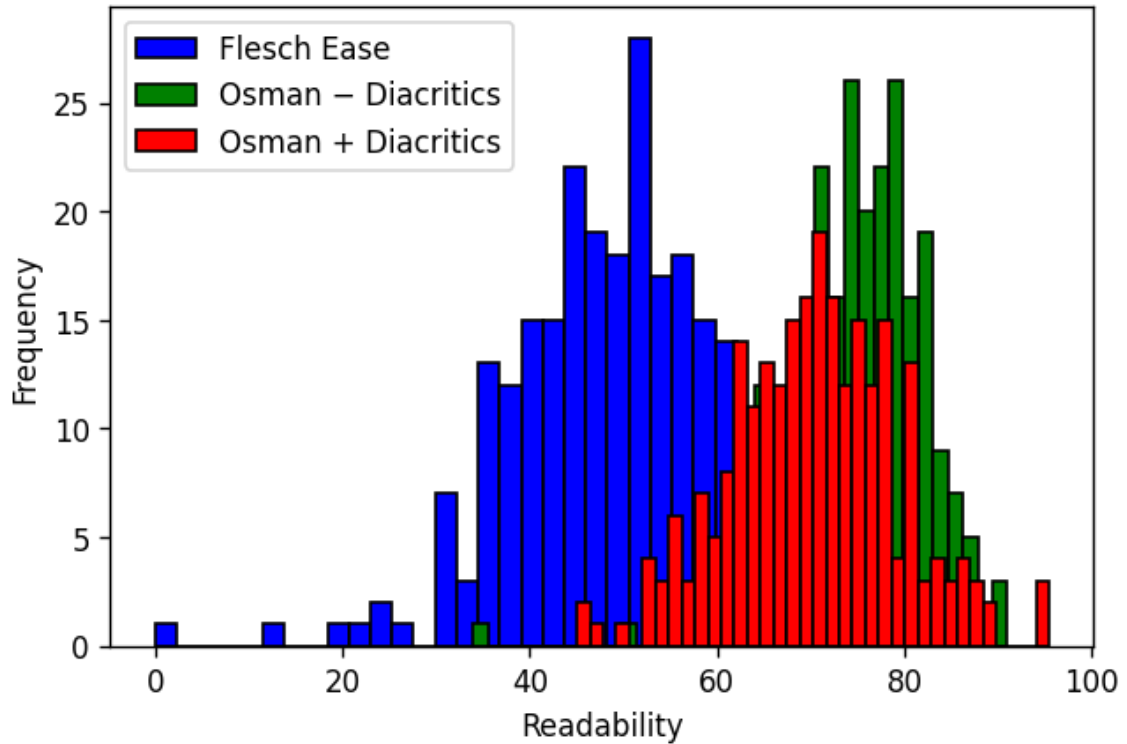


Figure 4.2: Histogram of readability comparisons between english and arabic with and without diacritics

4. Food
5. Health
6. Music
7. Science and Technology
8. Sports

The documents in the corpus are distributed as shown in Table[4.6] and Figure[4.4] with Science and technology ranking first between the other categories.

The general topic focus can be deduced from the previous 2 figures which is that scientific and technological topics are the dominants between the other categories.

For a summarization task, we need to ensure that entities are preserved between the original document (article) and its summary(abstract). So we applied entity recognition on each article and abstract, and computed the cosine similarity to have an idea of how much the entities' appearance in summaries is important. The entities that we have focused on were "Organization, Person, Location, Miscellaneous" and the obtained results show that there is a

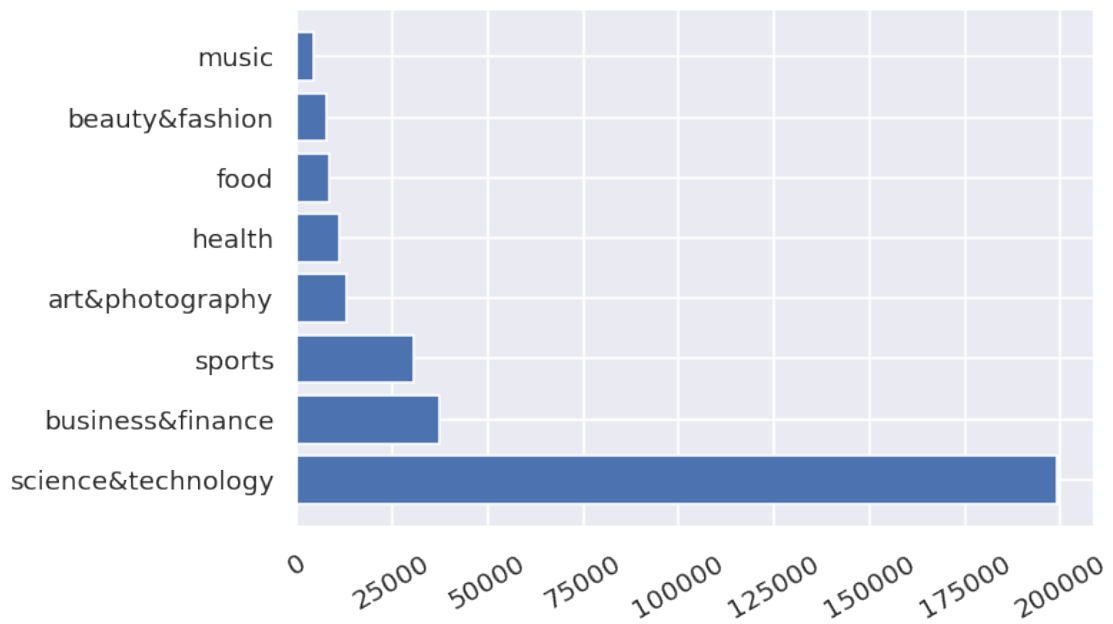


Figure 4.3: Bar chart showing the distribution of document labels per category

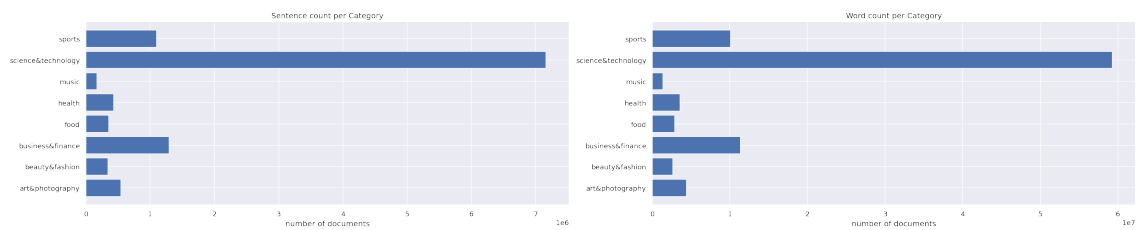


Figure 4.4: Bar chart showing the distribution of corpus sentences and words per category

Category	% Distribution
art/photography	4.17%
beauty/fashion	2.48%
business/finance	11.94%
food	2.73%
health	3.62%
music	1.46%
science/technology	63.76%
sports	9.83%

Table 4.6: Distribution of documents between different categories

high similarity on the skeletal level of the relation between articles and abstracts Figure[4.5]

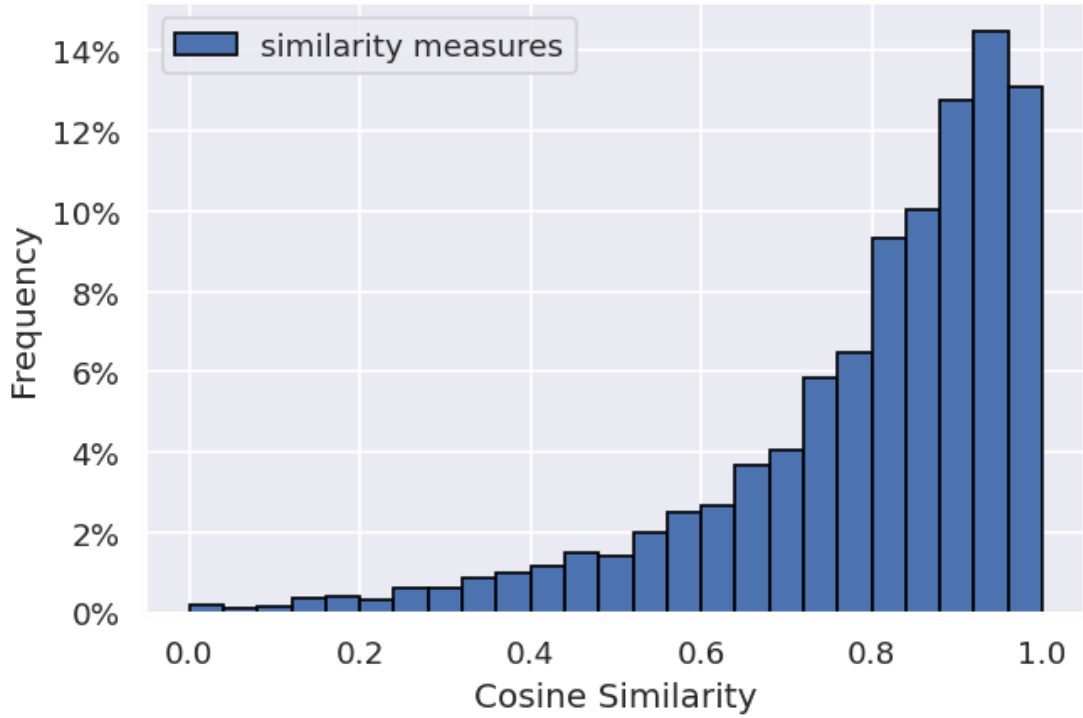


Figure 4.5: Cosine Similarity of named entity counts between abstracts and articles

4.1.4 *Kalimat Dataset*

We will evaluate the models trained on KALIMAT dataset [60], which is a multipurpose Arabic corpus Dataset that contains 20,291 Arabic articles collected from the Omani newspaper Alwatan. Mainly it is used for extractive summaries, and named entity recognition. The data has 6 categories: culture, economy, local-news, international-news, religion, and sports in modern standard arabic language.

4.2 Data Preprocessing

The raw documents in the translated corpus went through a pipeline of processes in order to extract the meta data required to highlight critical information that will be fed into the model later. Stanza [61] a multi-language NLP toolkit was used for tokenization and sentence segmentation. Figure[4.6] shows an example of how a passage gets split into tokens and each sentence consists of a group of tokens.

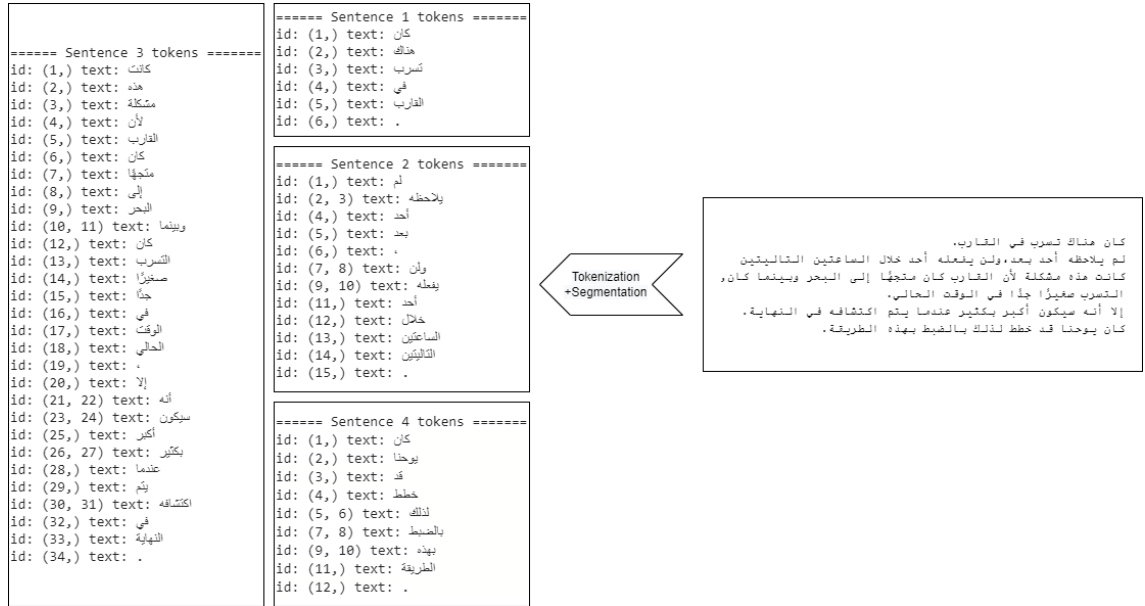


Figure 4.6: Sample representation of tokenizing and segmenting an arabic passage

An associated labeling vector with each document will hold the value 1 if the i^{th} sentence is an important sentence and 0 otherwise. This will be used later in the training of the models.

After splitting and labeling each sentence in the document. The text file samples are then converted into JSON files as it is a more robust data structure to work with.

```

1 [
2     { #For each document
3         "src": list of tokenized words from main article,
4         "tgt": list of tokenized words from the target extractive
5             summary,
6     },
7 ]

```

Finally the structured JSON files are converted into binary files to enhance the parallelism of the model training.

4.3 Training Details

The dataset have been split into train, test, and validation sets as shown in Table[4.7]

Dataset Split	Number of Documents per Set
Train Set	287,226
Test Set	11,489
Validation Set	13,367

Table 4.7: Distribution of documents between the split sets

Training was ran on an NVIDIA Tesla V100 32GB GPU, the pretrained bert model used was "arabertv02", and the following hyper-parameters were used on the 3 models trained:

We applied 50,000 training steps with a batch size of 1,000, optimization algorithm chosen was *Adam* with a learning rate $5 \cdot 10^{-5}$, $\text{beta_1} = 0.9$, $\text{beta_2} = 0.999$, $\text{epsilon} = 10^{-5}$, and dropout rate of 0.1. Noam decay scheme was applied with 8,000 warm-up steps. In order to fix the randomness a constant seed of 666 was used. Table[4.8] shows the special hyper-parameters of each encoder.

Encoder	Value	Description
Classifier	128	Number of hidden layers
RNN	768	Number of features in hidden state
	1	Number of recurrent layers
	True	Bidirectional
Transformer	768	Number of expected features
	4	Number of heads in multi-head attention models
	512	Dimension of feed forward network model
	2	Number of Intermediate Layers

Table 4.8: Hyper-parameters of different encoders

We used the Binary Cross Entropy (xent) for a loss function and each sentence will be labeled with a score between 0 and 1.

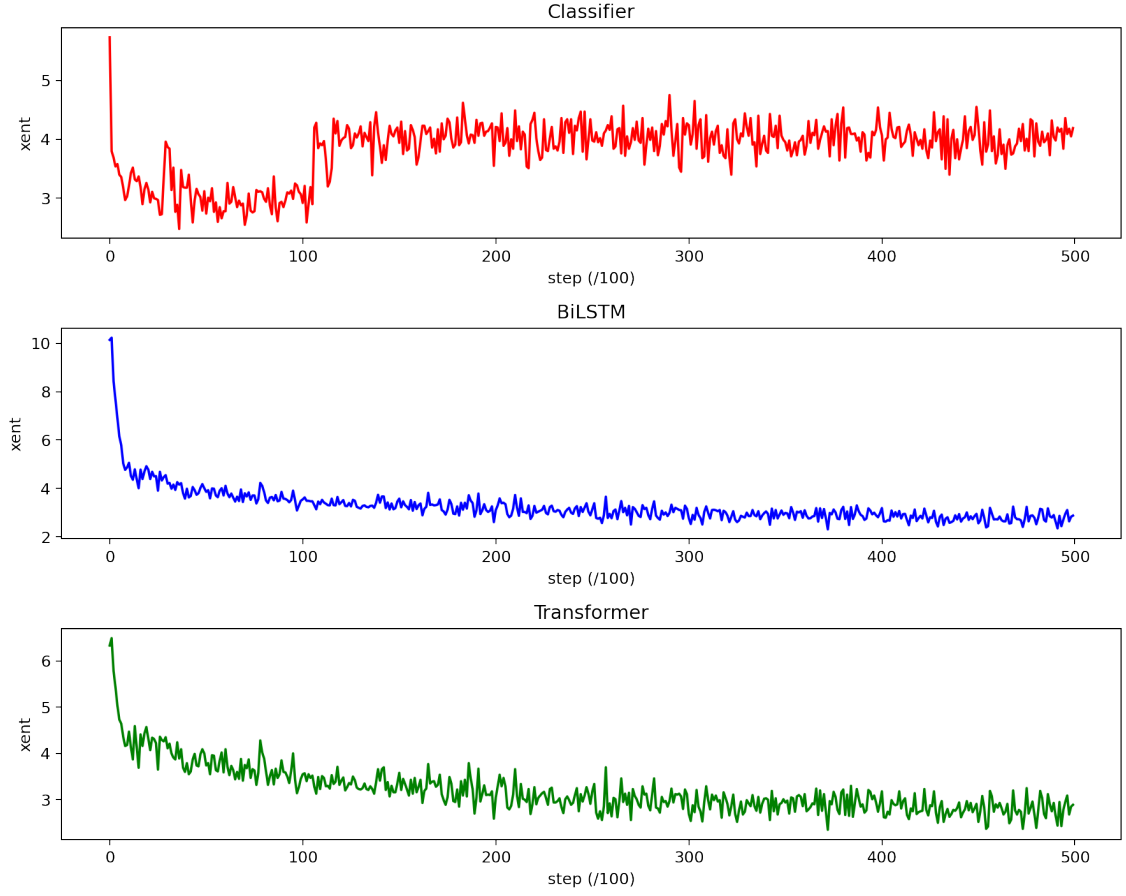


Figure 4.7: Loss training curve of each model

4.4 Results

Results are reported in Table[4.11]. The 3 metrics used for evaluation are BLEU (Average of the 4 BLEU evaluations over uni-,bi-,tri-, and quad-grams), ROUGE (Average between F1-Scores of ROUGE-1,ROUGE-2, and ROUGE-L), and the Siamese Similarity metric. Oracle Score represents the evaluation of human summaries that will be used for comparison with machine summaries. "SaraBert+RNN" was the best performing model based on the 3 evaluation metrics. Sample summaries given by the different models can be found in the Appendix[B].

Model	bleu_1	bleu_2	bleu_3	bleu_4	avg	min	max
SARaBert + Classifier	0.408	0.385	0.38	0.376	0.387	0.0	0.99
SARaBert + RNN	0.474	0.45	0.445	0.441	0.452	0.0	0.99
SARaBert + Transformer	0.391	0.368	0.364	0.36	0.371	0.0	0.99
AraBert Embeddings + K-Means	0.284	0.254	0.249	0.246	0.258	0.0	0.991
AraVec + K-Means	0.345	0.334	0.331	0.328	0.335	0.0	0.973
Bag of Words	0.361	0.351	0.348	0.345	0.352	0.0	0.98

Table 4.9: BLEU scores over Kalimat Dataset

Model	rouge_1	rouge_2	rouge_L	avg	min	max
SARaBert + Classifier	0.529	0.466	0.51	0.502	0.008	0.99
SARaBert + RNN	0.588	0.524	0.568	0.56	0.008	0.99
SARaBert + Transformer	0.512	0.444	0.491	0.482	0.017	0.99
AraBert Embeddings + K-Means	0.445	0.366	0.427	0.414	0.008	1.0
AraVec + K-Means	0.545	0.49	0.535	0.523	0.0	1.1
Bag of Words	0.47	0.415	0.46	0.448	0.0	1.1

Table 4.10: ROUGE scores over Kalimat Dataset

Model	BLEU	ROUGE	Siamese Similarity	BS	TS-SS
Oracle Score	0.532 \pm 0.27	0.699 \pm 0.23	0.326 \pm 0.3	0.83	0.003
SARaBert + Classifier	0.387 \pm 0.3	0.502 \pm 0.31	0.211 \pm 0.158	0.76	0.018
SARaBert + RNN	0.452 \pm 0.2	0.560 \pm 0.26	0.282 \pm 0.18	0.79	0.017
SARaBert + Transformer	0.371 \pm 0.3	0.482 \pm 0.31	0.240 \pm 0.17	0.77	0.019
AraBert Embeddings + K-Means	0.258 \pm 0.21	0.414 \pm 0.25	0.216 \pm 0.1	0.76	0.032
AraVec + K-Means	0.335 \pm 0.28	0.523 \pm 0.28	0.191 \pm 0.18	0.74	0.039
Bag of Words	0.352 \pm 0.35	0.448 \pm 0.36	0.217 \pm 0.28	0.76	0.111

Table 4.11: Results of different models under several evaluation metrics

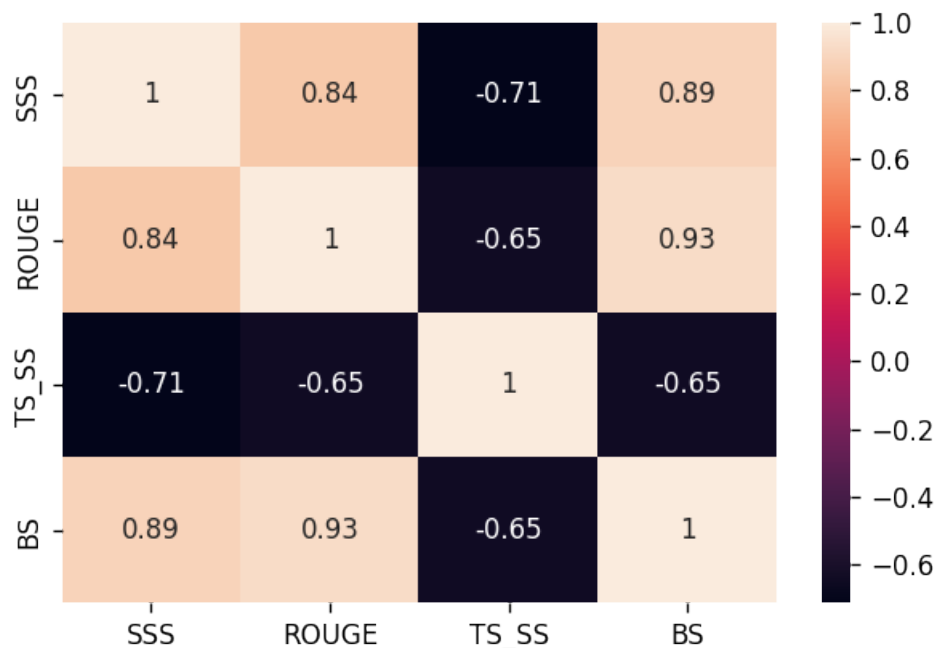


Figure 4.8: Pearson Correlation between the similarity metrics

4.5 Discussion

It appears that the usage of BiLSTM encoder did better than the transformer encoder, this could be due to the need of transformers for huge amount of data to train all its attention heads and layers. Moreover, the summarization layers are fed only the $[CLS]$ tokens making the size of the input equal to the number of sentences, which is on average 7, this might indicate that the hidden layers of the BiLSTM encoder are able to cover the small input size having a similar performance to the attention mechanism of the transformers. One problem remains is the incapability of feeding very large documents into the models to obtain a single global lookup on the document instead of segmenting the document and losing linked contexts between the trimmed passages. The translated documents fed into the model have shown no correlation between the readability variation and the ROUGE, meaning that translation did not affect the quality and content of the translated articles. However, the topic category and structure of the document had an effect on the results and showing that the model can best handle news articles. In addition, the short sentences have negatively affected the sentence scoring and produced low information coverage as shown in Figure[B.3] Future work should focus on enhancing SARA-Bert, the contextualized embeddings can be appended with regular embeddings such as GloVe, WordNet, FastText, etc. in order to see if the performance gets boosted by generating hybrid embeddings.

Chapter 5

Conclusion

Natural language understanding has shown to be a more complex task when dealing with the Arabic language in comparison with the English language. Text summarization is one of the important areas of research in NLP, as textual data keep increasing by day. Prior art have introduced statistical approaches and sequence-to-sequence machine learning models to generate text summaries. The proposed model "*SArBERT*" works on the summarization task for MSA NLP. We also created a similarity metric that evaluated the similarity of 2 documents on the semantic level and not just the syntactic level. We compared it with other types of similarity metrics that evaluate on the syntax-level, embeddings-level, and geometry-level. Experiments on KALIMAT corpus dataset have shown that *SArBERT* with RNN encoding and achieved the best performance on the syntax (ROUGE 0.56) and semantic (SSS 0.282). The model worked best with objective documents on contrary to subjective documents. The SSS metric highly correlates with other similarity metrics but requires to have a mechanism that can focus on context changing parameters such as a negation word. Future work should investigate the expansion of the model's understanding to other Arabic Dialects. As for the SSS, a modified version of cosine similarity should be implemented to allow certain dimensions to have higher or lower impact, because the cosine similarity by itself treats all dimensions equally.

Appendix A

Siamese Semantic Similarity Experiments

The metric that was discussed in Section [3.3](#) had to be tested for different cases to visualise and evaluate the overall performance of this metric.

We applied SSS on a set of candidates shown in Table [A.1](#) to the reference found at the first row of that table. The results of SSS and BertScore can be found in Table [A.2](#) along with the values required to compute SSS. The correlation between BertScore and SSS is 0.97 showing that both metrics behave very similar but SSS tends to be more strict with its evaluations.

ID	Text
0	أكل التلميذ سامي طعامه في المدرسة صباحا يوم الثلاثاء. طعامه كان تفاحة
1	هرب
2	أكل
3	أكل التلميذ سامي طعامه في المدرسة صباحا يوم الثلاثاء. طعامه كان تفاحة
4	تناول التلميذ سامي طعامه في المدرسة صباحا يوم الثلاثاء. طعامه كان تفاحة
5	تناول سامي طعامه في المدرسة صباحا يوم الثلاثاء. طعامه كان تفاحة
6	تناول سامي تفاحة في المدرسة صباح الثلاثاء
7	أكل سامي
8	أكل سامي تفاحة
9	أكل سامي دجاجة
10	أكل سامي طعاما
11	أكل التفاح سامي
12	أكلت سامية طعامها
13	في المدرسة تناول سامي وجبة تفاح
14	في المدرسة أكل سامي
15	نهار الثلاثاء صباحا أكل سامي تفاحة
16	فازت المثلة جيسيكا بجائزة اوسكار
17	كان يصوم معها حتى فهم أن ذلك يعني أنه لا يستطيع الأكل
18	اسمح لي أن أساعدك في أمتعتك
19	استيقظ سامي صباح يوم الثلاثاء متحمسا للذهاب الى المدرسة لانه يريد ان يتناول تفاحة هناك
20	أكل سامي تفاحة أكل سامي تفاحة أكل سامي تفاحة أكل سامي تفاحة أكل سامي تفاحة أكل أكل أكل أكل أكل أكل التلميذ سامي سامي
21	سامي طعامه المدرسة المدرسة في المدرسة المدرسة صباحا يوم الثلاثاء يوم يوم يوم. طعامه كان تفاحة تفاحة تفاحة تفاحة

Table A.1: List of a reference sentence (1) and set of candidate sentences(2 to 19)

ID	Cosine	ROUGE	norm₂	FD	SSS	BS
1	0.598	0.0	9.406	0.923	0.0	0.631
2	0.638	0.111	9.1	0.923	0.065	0.682
3	1.0	1.0	1.0	1.0	1.0	1.0
4	0.995	0.909	2.012	1.0	0.901	0.971
5	0.980	0.857	2.929	0.993	0.829	0.935
6	0.877	0.41	5.719	0.923	0.329	0.772
7	0.734	0.205	8.86	0.923	0.137	0.697
8	0.816	0.286	6.894	0.923	0.213	0.728
9	0.7630	0.19	7.789	0.923	0.133	0.693
10	0.8000	0.19	7.526	0.923	0.139	0.718
11	0.763	0.19	7.926	0.923	0.133	0.702
12	0.824	0.0	6.756	0.923	0.0	0.717
13	0.891	0.238	5.396	0.923	0.194	0.74
14	0.812	0.314	7.452	0.923	0.233	0.722
15	0.904	0.314	5.205	0.923	0.26	0.741
16	0.701	0.0	8.136	0.923	0.0	0.662
17	0.811	0.058	6.777	0.923	0.043	0.669
18	0.722	0.078	8.0	0.923	0.052	0.623
19	0.845	0.257	6.247	0.923	0.199	0.743
20	0.677	0.286	9.243	0.217	0.041	0.616
21	0.869	0.874	6.778	0.433	0.326	0.778

Table A.2: Results of computing SSS and BertScore (BS) along with other required computations for the candidates in Table[A.1] linked by ID

Appendix B

Sample Summaries form SaraBert

This section will present sample summaries extracted via SaraBert, highlights are used to visualize the selected top 3 sentences selected to be considered as the most informative sentences.

حددت وزارة الدفاع الروسية موعدًا نهائيًا آخر للجنود الأوكرانيين في مدينة ماريوبول الاستراتيجية بجنوب شرق البلاد للاستسلام ، قائلة إنه سيتم إنقاذ حياة الجنود داخل مصانع الصلب إذا أوقفوا ما أسمته "المقاومة غير المنطقية". وقالت الوزارة في بيان يوم الثلاثاء "كل من ألقى أسلحته مضمون للبقاء على قيد الحياة" مضيفة أن القوات ستكون قادرة على الانسحاب من مصنع الصلب بين الساعة 2 مساءً و 4 مساءً بتوقيت موسكو (11:00:13:00:00 بتوقيت جرينتش). "بدون استثناء ، بدون أسلحة وبلا ذخيرة". "ولم يتم الرد يوم الأحد على طلب مماثل وهو الاستسلام أو الموت ، بعد يوم من ادعاء الجيش الروسي أنه "ظهر بالكام" المدينة. وقال إدوارد باسورين المتحدث باسم الانفصاليين المدعومين من روسيا في منطقة دونباس يوم الثلاثاء إن الجماعات الهجومية تحركت إلى مصنع أزوفستال للصلب في محاولة لاقتلاع القوات الأوكرانية من جذورها. في وقت لاحق من اليوم ، قالت وزارة الدفاع إن القوات الروسية فتحت سمرًا إنسانيًا حتى تتمكن القوات الأوكرانية التي وافقت على إلقاء أسلحتها من مغادرة المدينة المحاصرة. وقالت الوزارة إن "القوات المسلحة الروسية فتحت سمرًا إنسانيًا لسحب أفراد الجيش الأوكراني الذين أقوا طواعية أسلحتهم وسلحي التشكيلات القومية" ، مضيفة أن السمر الأمن افتتح في الساعة 2 مساءً (11:00:00 بتوقيت جرينتش). وقال مجلس مدينة ماريوبول إن ما لا يقل عن 1000 مدني يختبئون في ملاجئ تحت مصنع الصلب الضخم الذي يحتوي على عدد لا يحصى من المبانى والأفران المتفجرة ومسارات السكك الحديدية. ونفت روسيا التقارير التي تفيد بوجود نساء وأطفال ومدنيين آخرين هناك ودعت كييف إلى "ممارسة العقل". وقالت وزارة الدفاع: "ندعو مرة أخرى سلطات كييف لإظهار السبب وإعطاء الأوامر للمقاتلين لوقف مقاومتهم المحققة". "لكن ، مع إدراك أنهم لن يتلقوا مثل هذه التعليمات والأوامر من سلطات كييف ، فإننا ندعو [المقاتلين] إلى اتخاذ هذا القرار طوعًا وإلقاء أسلحتهم". ناشد الرائد سيرهي فولينا ، قائد اللواء 36 من مشاة البحرية الأوكرانية ، والذي لا يزال يُقاتل في ماريوبول ، المساعدة يوم الاثنين في رسالة إلى البابا فرنسيس. "هذا ما يبدو عليه الجحيم على الأرض ... إبان الوقت [للحصول] على المساعدة ليس فقط بالصلاة. وقال في الرسالة ، وفقًا لمقتطفات نشرها سفير أوكرانيا في الفاتيكان على تويتر "إنقاذ حياتنا من أيدي الشيطان". وصف الرئيس الأوكراني فولوديمير زيلينسكي الوضع في ماريوبول بأنه "غير إنساني" وقال: "تحاول روسيا عمدًا تدمير كل من هناك". كما هدد بالانسحاب من محادثات السلام مع روسيا في حالة مقتل القوات الأوكرانية في المدينة. وقالت روسيا إن أوكرانيا فقدت أكثر من 4000 جندي في ماريوبول حتى يوم السبت. ومع ذلك ، تقول كييف إن إجمالي خسائر القوات على مستوى البلاد حتى الآن في الحرب ما بين 2500 و 1,300 أعلن زيلينسكي في وقت متأخر يوم الاثنين أن روسيا شنت هجومها الجديد المتوقع الذي يركز على منطقة دونباس بشرق أوكرانيا. أكد وزير الدفاع الروسي سيرجي شويغو أن قواته "تتخذ تدريجياً خطتنا لتحرير" شرق أوكرانيا. وقال في لقاء متلفز مع قادة عسكريين روس: "تتخذ إجراءات لإعادة الحياة السلمية". ركزت بعض أعنف المعارك في الحملة الروسية حول ماريوبول ، وهي مدينة ساحلية على بحر آزوف كانت مسرحًا لأعنف المعارك وأسوأ كارثة إنسانية في الحرب. حوصر عشرات الآلاف من السكان في ماريوبول من دون طعام أو ماء ، وتناثر الجثث في الشوارع. تعتقد أوكرانيا أن أكثر من 20 ألف مدني قتلوا هناك. سيؤدي الاستيلاء عليها إلى ربط الأراضي الانفصالية الموالية لروسيا بمنطقة القرم التي ضمتها موسكو في عام 2014 ، وتحرير القوات المحاصرة للهجوم في أماكن أخرى في دونباس.

Figure B.1: Sample SaraBert summarization. Yellow highlighting represents summary extraction using MLP as encoder, Green represents RNN encoder, and Blue represents Transformer

بدي منتجون ومثّلون عرب أسفهم من استمرار انتهاج التلفزيونات العربية سياسة عرض الاعمال المحلية العربية دون تبادل عرضها فيما بينها مشيرين إلى ان ذلك عائق في سبيل مزيد من التواصل بين شعوب المنطقة فيما لا توجد مشكلة في عرض الاعمال الاجنبية . طرح مراسل وكالة الانباء الالمانية القضية على عدد من هؤلاء الفنانين والمنتجين خلال فعاليات مهرجان القاهرة للأذاعة والتلفزيون العاشر ففي البداية أكد المنتج الاردني جواد مرقه رئيس اتحاد المنتجين العرب أن المشكلة قديمة ومستمرة فالبعض يقول عددي كفايتي من الاعمال ولست في حاجة إلى المزيد وهذا قد يكون حقهم لكن إذا كنا نتحدث عن سوق تجارية فإن المنطق يقول إنك لكي تبيع فلا بد أن تشتري ود المشتريين القضية الآن تحولت إلى إمكانيات مادية لدى المحطات . وتساءل قائلاً هل الاعلام العربي وخاصة المواد الدرامية لمجرد البيع والشراء فقط أم أنها للتبادل الثقافي بين الشعوب العربية والعزف على تنويعات عربية مختلفة ! إذا كان الحال كذلك فإن الدول العربية لابد أن تعرض كافة الاعمال العربية المتميزة لكن بعض الدول لا تقبل لهجات أخرى غير لهجتها إلى جانب الفصحى وأضاف جواد مرقه أن التلفزيونات العربية ما زالت تفكر تفكيراً قبطرياً والتفكير القبطري لا يصلح في هذه المرحلة لكن هذا هو الواقع ومن أجله كان إنشاء اتحاد المنتجين العرب لكنه استمرّد أن الاتحاد مثله مثل كل الاتحادات العربية يعاني لأنه لا يستطيع فرض رأيه على أحد عكس الاتحادات الماثلة في الخارج والتي تستطيع فرض الاسعار ومواعيد الدفع وعدد مرات العرض وغيره من الشروط والاخوة في المحطات العربية المختلفة يوقعون عقوداً معها ويلتزمون بها أما فيما بينهم فهناك الكثير من المعوقات نسعى دائماً لحلها ونأمل أن نصل إلى اتفاق سلمي يرضي الجميع . بينما ألقى المنتج والمخرج السوري مظهر الحكيم بالتعبية على المبدعين أنفسهم قائلاً المشكلة الاساسية أننا جميعاً نغلق ابوابنا ونوافذنا في وجه الشمس والشمس التي أقصدها هي حرية التعبير وحرية الرأي وحرية الاداء نحن مغلقون بسبب اتجاهاتنا الاعلامية والسياسية والاقتصادية وإن كان البعض يبدأ يتغلب على هذه المخاوف عن طريق إتاحة رأس مال حر نتج عنه شركات خاصة بدأت تعمل لكن أيضاً من منطلق ظرفها الخاصة وإمكانياتها المتاحة ويضيف لقد تحدثنا في كل المهرجانات العربية وطالبنا بفتح الابواب على جميع المستويات ولجميع الفنانين العرب أننا أحد المخرجين الذين تعاملوا مع مصريين وخليجيين ولبنانيين وسعوديين في أعالي لكن هل استطعت الوصول لا أقول للعالمية ولكن للعربية هذا لم يحدث لأن المحطات وقعت اتفاقات مع شركات دول أخرى ومع دول دون أخرى . وقال الممثل الأردني جميل عواد إن المشكلة خاصة بعرض الدراما التلفزيونية عبر الشاشات العربية فهناك دول تقصر شاشاتها على الدراما الخاصة بها فقط فكيف لنا أن نواجه الدراما العربية أو نصل إليها من حيث الإمكانيات والجودة والانتشار وأضاف أنه ليس هناك معايير لشراء الدراما فلا نعرف هل المطلوب هو الجودة أم التكاليف وإلى متى يتم الشراء بالمتنر والكيلو هذه إشكاليات لا بد من مناقشتها واتخاذ قوانين خاصة بها وعلى اتحاد المنتجين العرب متابعة التنفيذ . أما المنتج المصري إبراهيم أبو ذكري فألقى باللوم على تحول بعض المنتجين إلى مسامرة إنتاج مؤكداً أننا أصبحنا نفقد المنتج الحقيقي الواعي وأشار إلى أنه مع انتشار الدراما العربية في مختلف الدول العربية ومع استمرار استخدام اللهجات المختلفة حتى يتعرف العرب على لهجات أشقائهم العرب وهذا سبيل الوحدة والتواصل وأضاف أبو ذكري أن رفض استخدام الدوبلاج لأنه يقتل الدراما العربية ويحرماننا من متعة معرفة اللهجات هذه هي المتعة الحقيقية أنا وأنت تماماً أن الامر يتوقف على الوقت فقط للتعود على اللهجات المختلفة بدليل فهم كل العرب للهجة المصرية التي تنقسم بدورها إلى عدد من اللهجات مثل العامية والصعيدية والريفية وغيرها دون أي مشكلة لتعودهم عليها

Figure B.2: Sample SaraBert+BiLSTM of a good summarization

أسمن النظر في عينيه غاص صيفاً حتى وصل إليها لمعت عينها ببريق ومض للحظات ثم استكانت لموت متأصل ينظر حواله المكان غارق في عتمة كئيبة اتجه ناحية الستائر جذبها ليسبح مجالاً للنور والهواء بالدخول فجاء الحائط الأصم ينف صلباً وراءها بغياء مثير للحلق انتابته شعيرة مؤلمة عاد والقرب منها من جديد جثا على ركبتيه أمامها بكى وجهه بالحنن المتدفق مع نبضه أسك بكها فرد أصابعه المنتشحة قبل راحتها بتودد رحيم خرج صوته يرن بلحن التعاطف بسأله :

سأنا جاء بك إلى هنا!

لم ترد عليه رغم أنها كانت تنظر إلى وجهه مباشرة

رايت صورتك كان يجب أن أدخل إليك لا تخافي لست كالأخرين أردت أن أفهم أذا هو العمل الذي أخبرتني به ذاك اليوم ! ألم تقولي أنه كان وظيفة في مجال السياحة ! هل كنت على علم بنوعية العمل ! كان صوته يغلغه الرجاء بالحاج لكنها لم تحرك ساكناً كانت مثل هذا الجدار الأصم وراء الستائر المحلية هل كنت على علم ! هه أجيبيني ! هل كنت على علم ! كان يزحف إلى صوته شيء من العصب ربما لخصتها الذي أثار فيه مشاعر متضاربة ويعنف كانت عينها تلعب للحظات ثم ما لبثت أن تخب وتظفي أسسها من كثفها هزها : من الواضح لي أنك في سجن هل أنت سجنانة 1 أجيبني أرجوك ... تقاطعت فطرات دمع في عينها حتى جمعت سحابة بدأت تقبض ثم تنهس وجسدها كله ينتفض بين ذراعيه اللتين أحاطتا بها ورأسها على صدره يثبت متصلة العضلات لم تزل لم تستسلم لآلامه المطروح بسخاء بقيت مشتتة بالمها وبأسها وخوفها زحسها بقوة أكبر هاسا : لا تخافي أرجوك لست هذا لأي عرض غير أن أراك وأعرف منك لماذا أنت هنا ! وكيف بإمكان مساعذك أرجوك تحول بكائها إلى نحيب ثم تشيح النطق معه فزاده أكثر أرجوك اهادي ! ! ... أرجوك أرجوك ! ! ... وبعد عاصفة صارخة موجهة بدات تهدأ وتستكين ثم بدات ترد على أسئلته يرقع صوتها شهقات تخرج من أعصافها عليها تخرج معها كل مخزون القهر والألال الذي تعرضت له لم تكن فصتها بالعربية تماماً عليه وربما توقعها منذ أول مرة صرحت له هناك في المطعم الصغير في بلدها ... أراد أن يحذر ها لكنه فؤوم ظنونه وحزنه يته الحذاء أذاك وصمتاً لمأذا صمت ! لم يشعر في حياته بالكبر من القدم الذي يشعر به الآن وهل كان يتوقع أن يكون لكلامه حينها أثر فعال ! لم يكن متكادبل كان شبه مناكس من أنه خارج إطار الصورة ليس له يد في رسم أحداثها لكن لم لم يحاول شيئاً ما ! ! وما شعر أنه لين معنى بالأمر وأن تدخله سيكون سخيلاً وما كلمة ! ! لعلها نلعت أذاك ! ! أحس بالإضافة إلى كل المشاعر المتنافضة المؤلمة بتنافسها بالم الإحساس بالذنب لئلا يشك خفي لم يعان عن نفسه بوضوح ولكنه جثم على قلبه مثل غراب ينقع إنها تجارة شبكة مثل شبكة العنكبوت بدقتها وخبتها وإحكامها ونشاقفيتها ومثل شبكة الصيد بقوتها بين أين تبدأ ! ! وأين تنتهي ! ! من يحرك خيوطها ! ! من يذري ! ! احتوى جسدها الصغير المنتهك في صدره وعدها بأن يحرقها بنظرت إليه برجاء يأس ..

-أيتها الصغيرة ماذا تعرضت له حتى وصلت لهذا الحال ! قصت عليه نفاقاً مشتردة من قصص ليل شتى وحبوات متنوعة متفاوتة بالتراسة والدناءة والبهيمية لا لا لا لا مستحيل ماذا بقي منك أيتها الصغيرة على هذا الحال لم يتوقع أن ينفذ غير حطام شوته عاصفة هوجاء ولكنه قرر بإصرار ضغط على الجرس سمع صوت المفاح في الباب أطل الرجل الذي قدم به بوجه محايد أثار فيه الرغبة في لكحه حتى يتورم ويذرف دال له : أريدها أن تخرج معي ! ! وأشار إليها برأسه أجابه بابتكيزية المتكبرة المثيرة للاشمئزاز

-لا لا لا لا هذا غير مسموح بإمكانك التحدث مع الإدارة يا سيدي . .

-الفتك إليها عاد ليجلس إلى جانبها على السرير لفها من كثفها بنراعيه همس : لا تخافي سأسعى جاهداً إن أعيب طويلاً همست له بصوت أخذ بما بقي فيه من تجلد .

Figure B.3: Sample SaraBert+BiLSTM of a bad summarization

From the summaries in Figure[B.2] and Figure[B.3] it can be seen that the model best works with documents are have the type of journalism and not.

Appendix C

Translation Quality

In this section we will present sample translations and discuss the quality and performance effect on the results.

A Georgia man is heading to prison for nearly 10 years for his role in a credit card fraud gang associated with a website called Carder.su that is linked to \$50 million in worldwide losses. Cameron Harrison, aka "Kilobit," 28, of Augusta, Georgia, pleaded guilty to possessing more than 260 compromised credit and debit card numbers, which were recovered from his computer and email accounts following his arrest. He also admitted, according to the Department of Justice, that the Carder.su organization committed money laundering, narcotics trafficking and computer crimes. He said members tried to avoid detection by communicating through various encrypted forums, such as chat rooms, private messaging systems and virtual private networks. Harrison was identified when he purchased a counterfeit Georgia driver's license from an undercover special agent through the Carder.su network. During interactions with the undercover special agent, Harrison admitted he had been a vendor of counterfeit identifications in the defunct cyberfraud organization "ShadowCrew." Fifty-five individuals were charged in four separate indictments in Operation Open Market, run by Homeland Security Investigations and the U.S. Secret Service. To date, 26 individuals have been convicted and the rest are either fugitives or are awaiting trial. Harrison pleaded guilty to participating in a racketeer influenced corrupt organization, conspiracy to engage in a racketeer influenced and corrupt organization, and trafficking in and production of false identification documents. "This significant sentence is entirely fitting given that this defendant's actions and those of the larger criminal organization harmed countless innocent Americans and seriously compromised our financial system," said Homeland Security Investigations Executive Associate Director Peter T. Edge. "Criminals like this defendant who believe they can elude detection by hiding behind their computer screens here and overseas are discovering that cyberspace affords no refuge from American justice."

Figure C.1: Sample English passage with Fleach score of 11.38 (Hard to read)

يتوجه رجل جورجيا إلى السجن لمدة 10 سنوات تقريبا لدوره في عصابة احتيال ببطاقة الائتمان المرتبطة بموقع إلكتروني يسمى Carder.SU المرتبط بمبلغ 50 مليون دولار في الخسائر في جميع أنحاء العالم. أقر كامبيرون هاريسون، المعروف أيضا باسم "كلوبيت"، 28، من أوغستا، جورجيا، بالذنب على امتلاك أكثر من 260 أرقام بطاقات ائتمان وخصم الخصم، والتي تم استردادها من أجهزة الكمبيوتر الخاصة به وحسابات البريد الإلكتروني بعد اعتقاله. كما اعترف، وفقا لوزارة العدل، أن منظمة الكاردر. ارتكبت غسل الأموال واتجار المخدرات وجرائم الكمبيوتر. وقال إن الأعضاء حاولوا تجنب الكشف عن طريق الاتصال من خلال مختلف المنتديات المشفرة، مثل غرف الدردشة وأنظمة المراسلة الخاصة والشبكات الخاصة الافتراضية. تم تحديد هاريسون عندما اشترى رخصة قيادة جورجيا المزيفة من وكيل خاص سري من خلال شبكة Carder.Su أثناء التفاعلات مع الوكيل الخاص السري، اعترف هاريسون بأنه بائع للتزديدات المزيفة في منظمة سيثيرفراود بتفين. "Shadowcrew" تم اتهام خمسة وخمسين فردا في أربع لوائح اتهام منفصلة في عملية السوق المفتوحة، وتشغيلها من خلال تحقيقات الأمن الداخلي والخدمات السرية الأمريكية. حتى الآن، تم إدانة 26 شخصا والباقي إما الهاربين أو ينتظرون المحاكمة. أقر هاريسون بأنه مذنب للمشاركة في مضرب يؤثر على المنظمة الفاسدة، والتأمر للانخراط في مضرب يتأثر ونظام فاسد، والاتجار بمستندات الهوية الخاطئة وإنتاجها. وقال مدير الأمن الداخلي للسيارات التنفيذية بيتر تي إيدج "هذه الجملة الهامة تتوافق تماما بالنظر إلى أن أفعال المدعى عليه هذه وأكبر من المنظمة الإجرامية الأكبر التي أضرت عدد لا يحصى من الأميركيين الأبرياء الذين لا حصر لهم الذين يعانون من نظامنا المالي." "مجرمو مثل هذا المدعى عليه الذين يعتقدون أنهم يستطيعون الاكتشاف عن طريق الاختباء وراء شاشات الكمبيوتر هنا والخارج لا يكتشفون أن الفضاء الإلكتروني لا يملأ من العدالة الأمريكية."

Figure C.2: Sample Arabic passage (translation of Figure[C.1]) with Osman score of 66.73 (Slight hard to read)

When Grandma Nelly lays down a challenge, you'd better accept it. And she wasn't shooting low. Nelly was gunning for NBA star Dwyane Wade. "Dwyane Wade," she says, pointing her finger at the camera in a YouTube video. "On my 90th birthday, I want to play one-(on)-one with you." On Tuesday, she got her wish. Grandma Nelly -- AKA Illuminada Magtoto -- met Wade at the Heat practice facility for a little shoot-around. "I want to play with you," said the grinning grandma who barely came up chest-high to the 6-foot, 4-inch guard. "Now I'm 90." Wade appeared to be just as touched by the encounter. "This gives me life. This gives me a purpose," he said after it was over. Wade sealed the date with a kiss on the hand. "Oh my God," she squealed. "I'm very, very happy. I'm very grateful," Nelly said. "This is my dream come true."

Figure C.3: Sample English passage with Fleach score of 83.69 (Easy to read)

عندما تضع الجدة نيللي تحديا، من الأفضل قبولها. وكانت لا تطلق النار منخفضة. نيللي كان يطبخ ل NBA Star Dwyane Wade. "Dwyane Wade"، تقول، تشير إلى إصبعها على الكاميرا في فيديو يوتيوب. "في عيد ميلادي التسعين، أريد أن ألعب (on) - One - معك." يوم الثلاثاء، حصلت على رغبته. Grandma Nelly - AKA Illuminada Magtoto - التقى واد في منشأة الممارسة الحرارية لقليل تبادل لإطلاق النار حولها. "أريد أن ألعب معك." "الآن أنا 90." بدا أن واد أن تكون مؤكدا تماما. وقال بعد انتهاء الأمر "هذا يعطيني الحياة. هذا يعطيني غرضا". أغلق واد التاريخ مع قبله في اليد. "يا إلهي". وقال نيللي "أنا سعيد جدا جدا. أنا ممتن للغاية". "هذا حلمي حقيقة".

Figure C.4: Sample Arabic passage (translation of Figure[C.3]) with Osman score of 88.01 (Easy to read)

The translations have shown no effect on the trained model, where the readability variation had no link with ROUGE results. We sampled 1000 documents and computed their respective readability metrics in Arabic and English (Osman and Flesch) along with the ROUGE value of the generated summary. By looking at table[C.2], it can be observed that no readability metric highly affects the ROUGE evaluation.

Index	Flesch	Osman	Rouge-AR	Readability-Diff
Flesch	1.0	0.76	-0.12	-0.64
Osman	0.76	1.0	-0.04	0.01
rouge-AR	-0.12	-0.04	1.0	0.14
Readability-Diff	-0.64	0.01	0.14	1.0

Table C.2: Correlation between the different features

Flesch	59.93 ± 10.13
Osman	74.88 ± 7.79
Rouge-AR	0.41 ± 0.14
Readability-Diff	14.95 ± 6.55

Table C.1: Average value for documents computed over 1000 sample documents

Bibliography

- [1] H. Jing, "Using hidden markov modeling to decompose human-written summaries," *Computational linguistics*, vol. 28, no. 4, pp. 527–543, 2002.
- [2] A. B. Al-Saleh and M. E. B. Menai, "Automatic arabic text summarization: A survey," *Artificial Intelligence Review*, vol. 45, no. 2, pp. 203–234, 2016.
- [3] K. S. Al Harazin, "Multi-document arabic text summarization," 2015.
- [4] K. Shaalan, "Nizar y. habash, introduction to arabic natural language processing (synthesis lectures on human language technologies)," *Machine Translation*, vol. 24, pp. 285–289, Dec. 2010. doi: [10.1007/s10590-011-9087-8](https://doi.org/10.1007/s10590-011-9087-8).
- [5] K. SHAALAN, M. MAGDY, and A. FAHMY, "Analysis and feedback of erroneous arabic verbs," *Natural Language Engineering*, vol. 21, no. 2, pp. 271–323, 2015. doi: [10.1017/S1351324913000223](https://doi.org/10.1017/S1351324913000223).
- [6] S. Izwaini, "Problems of arabic machine translation: Evaluation of three systems," in *Proceedings of the International Conference "The Challenge of Arabic for NLP/MT". 23 October 2006, London, United Kingdom. Pages*, 2006, pp. 118–148.
- [7] S. K. Ray and K. Shaalan, "A review and future perspectives of arabic question answering systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3169–3190, 2016. doi: [10.1109/TKDE.2016.2607201](https://doi.org/10.1109/TKDE.2016.2607201).
- [8] M. Korayem, D. Crandall, and M. Abdul-Mageed, "Subjectivity and sentiment analysis of arabic: A survey," in *Advanced Machine Learning Technologies and Applications*, A. E. Hassanien, A.-B. M. Salem, R. Ramadan, and T.-h. Kim, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 128–139, isbn: 978-3-642-35326-0.
- [9] N. Habash and O. Rambow, "MAGEAD: A morphological analyzer and generator for the Arabic dialects," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia: Association for Computational Linguistics, Jul. 2006, pp. 681–688. doi:

- 10.3115/1220175.1220261. [Online]. Available: <https://aclanthology.org/P06-1086>.
- [10] K. Shaalan, "A Survey of Arabic Named Entity Recognition and Classification," *Computational Linguistics*, vol. 40, no. 2, pp. 469–510, Jun. 2014, issn: 0891-2017. doi: 10.1162/COLI_a_00178. eprint: https://direct.mit.edu/coli/article-pdf/40/2/469/1803591/coli_a_00178.pdf. [Online]. Available: https://doi.org/10.1162/COLI%5C_a%5C_00178.
 - [11] P. T. Daniels, "The arabic writing system," *The Oxford handbook of Arabic linguistics*, pp. 422–431, 2013.
 - [12] K. C. Ryding, *A reference grammar of modern standard Arabic*. Cambridge university press, 2005.
 - [13] I. A. Al-Sughaiyer and I. A. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," *Journal of the American society for information science and technology*, vol. 55, no. 3, pp. 189–213, 2004.
 - [14] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, pp. 1–22, 2009.
 - [15] N. Y. Habash, "Introduction to arabic natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–187, 2010.
 - [16] R. Belkebir and A. Guessoum, "A supervised approach to arabic text summarization using adaboost," in *New contributions in information systems and technologies*, Springer, 2015, pp. 227–236.
 - [17] I. Imam, N. Nounou, A. Hamouda, and H. A. A. Khalek, "An ontology-based summarization system for arabic documents (ossad)," *International Journal of Computer Applications*, vol. 74, no. 17, pp. 38–43, 2013.
 - [18] K. Knight and D. Marcu, "Summarization beyond sentence extraction: A probabilistic approach to sentence compression," *Artificial Intelligence*, vol. 139, no. 1, pp. 91–107, 2002.
 - [19] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer: A parse-and-trim approach to headline generation," MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES, Tech. Rep., 2003.
 - [20] T. Cohn and M. Lapata, "Sentence compression beyond word deletion," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008, pp. 137–144.
 - [21] K. Woodsend, Y. Feng, and M. Lapata, "Generation with quasi-synchronous grammar," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010, pp. 513–523.

- [22] V. Patil, M. Krishnamoorthy, P. Oke, and M. Kiruthika, "A statistical approach for document summarization," *Department of Computer Engineering Fr. C. Rodrigues Institute of Technology, Vashi, Navi Mumbai, Maharashtra, India*, 2004.
- [23] F. Alotaiby, S. Foda, and I. Alkharashi, "New approaches to automatic headline generation for arabic documents," *Journal of Engineering and Computer Innovations*, vol. 3, no. 1, pp. 11–25, 2012.
- [24] R. Ferreira, L. de Souza Cabral, R. D. Lins, et al., "Assessing sentence scoring techniques for extractive text summarization," *Expert systems with applications*, vol. 40, no. 14, pp. 5755–5764, 2013.
- [25] Q. Al-Radaideh and M. Afif, "Arabic text summarization using aggregate similarity," in *International Arab conference on information technology (ACIT2009), Yemen*, 2009.
- [26] A. Haboush, M. Al-Zoubi, A. Momani, and M. Tarazi, "Arabic text summarization model using clustering techniques," *World of Computer Science and Information Technology Journal (WCSIT) ISSN*, pp. 2221–0741, 2012.
- [27] F. El-Ghannam and T. El-Shishtawy, "Multi-topic multi-document summarizer," *arXiv preprint arXiv:1401.0640*, 2014.
- [28] H. N. Fejer and N. Omar, "Automatic arabic text summarization using clustering and keyphrase extraction," in *Proceedings of the 6th International Conference on Information Technology and Multimedia*, IEEE, 2014, pp. 293–298.
- [29] N. M. Hewahi and K. A. Kwaik, "Automatic arabic text summarization system (aatss) based on semantic features extraction," *International Journal of Technology Diffusion (IJTD)*, vol. 3, no. 2, pp. 12–27, 2012.
- [30] M. El-Haj, U. Kruschwitz, and C. Fox, "Multi-document arabic text summarisation," in *2011 3rd Computer Science and Electronic Engineering Conference (CEECE)*, IEEE, 2011, pp. 40–44.
- [31] D. Miller, "Leveraging bert for extractive text summarization on lectures," *arXiv preprint arXiv:1906.04165*, 2019.
- [32] M. A. Fattah and F. Ren, "Ga, mr, ffnn, pnn and gmm based models for automatic text summarization," *Computer Speech & Language*, vol. 23, no. 1, pp. 126–144, 2009.
- [33] Q. A. Al-Radaideh and D. Q. Bataineh, "A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms," *Cognitive Computation*, vol. 10, no. 4, pp. 651–669, 2018.
- [34] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining text data*, Springer, 2012, pp. 43–76.

- [35] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [36] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [37] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.
- [38] R. Nallapati, B. Zhou, and M. Ma, "Classify or select: Neural architectures for extractive document summarization," *arXiv preprint arXiv:1611.04244*, 2016.
- [39] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 93–98.
- [40] M. Al-Maleh and S. Desouki, "Arabic text summarization using deep learning approach," *Journal of Big Data*, vol. 7, no. 1, pp. 1–17, 2020.
- [41] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.
- [42] A. M. Abu Nada, E. Alajrami, A. A. Al-Saqqa, and S. S. Abu-Naser, "Arabic text summarization using arabert model using extractive text summarization approach," 2020.
- [43] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.
- [44] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, *The interplay of variant, size, and task type in arabic pre-trained language models*, 2021. arXiv: [2103.06678 \[cs.CL\]](https://arxiv.org/abs/2103.06678).
- [45] N. Zalmout and N. Habash, "Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for arabic," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 704–713.
- [46] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.