Sentiment Analysis for Arabic Social media Movie Reviews Using Deep Learning

استخدام التعلم العميق لتحليل المشاعر لمراجعات الأفلام العربية في
مواقع التواصل الاجتماعي

by

FATEMA HAMAD MEZAHEM

Dissertation submitted in partial fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

October 2022

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the contents of this dissertation for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

This work is to apply sentiment analysis SA for Arabic movie reviews on social media. Automatically detecting attitude or sentiment in a text is often helpful. By classifying the data into positive, negative, or neutral emotions, SA aids in our understanding of the precise emotions that underlie the broader feelings that are typically associated with behavior. By utilizing the power of multiple word representations and deep learning approaches, this work seeks to enhance categorization performance. Through the use of mobile apps, the internet, and social media portals, there has been a tremendous increase of data in recent years. People are now able to share their opinions about specific topics because to the rapid development of technologies and social media platforms. People all around the world use a number of these social media sites frequently to share their evaluations and opinions of movies. By evaluating prior evaluations, it has become simpler for individuals to identify movies that live up to their expectations thanks to technologies like machine learning (ML) and deep learning (DL). Massive data can be collected every day from social media network such as YouTube, twitter, Instagram, and many other platforms. The tools used for collecting data are Vicinitas for Twitter and IGCommentExport for Instagram. The testing datasets were collected from mainly from Instagram for two Arabic movies reviews. The two movies are Wahed Tani which translates to (someone else) and Amahom which translate to (their uncle), Three datasets were employed, and several categorization models were compared across them. Prior to performing sentiment analysis, it is necessary to prepare the data so that it may be used to train machine learning (ML) algorithms. In order to label the data that was gathered from a corpus collection for ML use, manual annotation was made. For sentiment analysis, pre-processing is a crucial step in the data preparation process. Data pre-processing is a crucial step in NLP activities to enhance dataset performance and guarantee the accuracy of the emotive analysis. We translated some of the most common emojis as per its meaning in Arabic. There are different types of Arabic and the three main are Classical Arabic (CA), Modern standard Arabic (MSA), and Dialect Arabic language (DA). In this paper we are focusing on DA Arabic since it is commonly used on social media The main dataset was the Arabic Sentiment Analysis Dataset (ASAD) which presented a novel large Twitter-based benchmark (Alharbi et al., 2020). The proposed CNN, RNN, CNN-RNN, and BERT models were used in conjunction with the three datasets. With the Bert model

and in comparison, to the other examined models, two of these datasets were used. We test the CNN model first, then the LSTM, and finally the CNN-LSTM combo. After comparing these three modes, the best mode was chosen in order to compare it to the BERT model. The results of the hybrid CNN-LSTM model showed an accuracy of 90%. Finally, we compared CNN-LSTM with the BERT model Therefore, the BERT model outperformed all other classifiers in terms of accuracy (91%), recall (71%%), precision (83%), and F-measure (77%).

ملخص

هذا العمل لتطبيق قراءة المشاعر لمراجعات المستخدمين في مواقع التواصل الاجتماعي. غالبًا ما يكون الكشف التلقائي عن المواقف أو المشاعر في النص مفيدًا. من خلال تصنيف البيانات إلى مشاعر إيجابية أو سلبية أو محايدة ، يساعد SA في فهمنا للعواطف الدقيقة التي تكمن وراء المشاعر الأوسع التي ترتبط عادةً بالسلوك. من خلال الاستفادة من قوة تمثيلات الكلمات المتعددة وأساليب التعلم العميق ، يسعى هذا العمل إلى تحسين أداء التصنيف. من خلال استخدام تطبيقات الأجهزة المحمولة والإنترنت وبوابات الوسائط الاجتماعية ، كانت هناك زيادة هائلة في البيانات في السنوات الأخيرة. أصبح الناس الآن قادرين على مشاركة آرائهم حول مواضيع محددة بسبب التطور السريع للتقنيات ومنصات التواصل الاجتماعي. يستخدم الأشخاص في جميع أنحاء العالم عددًا من مواقع التواصل الاجتماعي هذه بشكل متكرر لمشاركة تقييماتهم وآرائهم حول الأفلام. يستخدم الأشخاص في جميع أنحاء العالم عددًا من مواقع الوسائط الاجتماعية هذه مجانًا من خلال تقييم التقييمات السابقة ، أصبح من الأسهل للأفراد تحديد الأفلام التي ترقى إلى مستوى توقعاتهم بفضل تقنيات مثل التعلم الآلي (ML) والتعلم العميق.(DL) . يمكن جمع بيانات ضخمة كل يوم من شبكات التواصل الاجتماعي مثل يوتيوب وتويتر وإنستغرام والعديد من المنصات الأخرى. الأدوات المستخدمة لجمع البيانات هي Vicinitas لمنصة تويتر و IGCommentExport لمنصة انستغرام تم جمع مجموعات بيانات الاختبار بشكل أساسي من الانستغرام لمراجعين عن فيلمين عربيين هما الفيلمين واحد تاني و عمهم. تم استخدام ثلاث مجموعات بيانات ، وتمت مقارنة العديد من نماذج التصنيف عبرها. قبل إجراء تحليل المشاعر ، من الضروري إعداد البيانات بحيث يمكن استخدامها لتدريب خوارزميات التعلم الآلي.(ML) . من أجل تسمية البيانات التي تم جمعها من مجموعة المراجع لاستخدام ML ، تم عمل التعليقات التوضيحية اليدوية. لتحليل المشاعر ، تعد المعالجة المسبقة خطوة حاسمة في عملية إعداد البيانات. تعد المعالجة المسبقة للبيانات خطوة حاسمة في أنشطة البرمجة اللغوية العصبية لتحسين أداء مجموعة البيانات وضمان دقة التحليل الانفعالي. قمنا بترجمة بعض الرموز التعبيرية الأكثر شيوعًا وفقًا لمعناها إلى اللغة العربية. هناك أنواع مختلفة من اللغة العربية وثلاثة أنواع رئيسية هي العربية الفصحى (CA) والعربية الفصحى الحديثة (MSA) واللغة العربية اللهجة.(DA) في هذه الورقة ، نركز على DA Arabic نظرًا لاستخدامه بشكل شائع على وسائل التواصل الاجتماعي. وكانت مجموعة البيانات الرئيسية هي مجموعة بيانات تحليل المشاعر العربية (ASAD)التي قدمت معيارًا جديدًا كبيرًا يعتمد على تويتر. تم استخدام نماذج CNN و RNN و CNN-RNN و BERT المقترحة بالاقتران مع مجموعات البيانات الثلاث. مع نموذج بيرت وبالمقارنة مع النماذج الأخرى التي تم فحصها ، تم استخدام مجموعتين من مجموعات البيانات هذه. نختبر نموذج CNN أولاً ، ثم LSTM ، وأخيراً مجموعة .CNN-LSTM . بعد مقارنة هذه الأوضاع الثلاثة ، تم اختيار أفضل وضع لمقارنته بنموذج.BERT أظهرت نتائج نموذج CNN-LSTM الهجين دقة 90٪. أخيرًا ، قمنا بمقارنة CNN-LSTM بنموذج BERT ، لذلك تفوق نموذج BERT على جميع المصنفات الأخرى من حيث الدقة (91٪) ، والتذكر (71 ٪) ، والدقة (83٪) ، وقياس. .(77٪) F .

ACKNOWLEDGEMENTS

I would like acknowledge and give gratitude to all British University in Dubai (BUiD) community the for giving me the chance to experiment and study informatics major that provided me with great knowledge.

I also want to express my gratitude to Dr. Manar Alkhatib for supervising this dissertation and making this work possible. I also thank Dr. Khaled Shaalan offering his support and advice during this research. Without their assistance, it would have been difficult to produce a research paper of this quality.

Finally, many thanks to my family, my parents, my husband and my sisters who were all motivating and supportive to my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 introduction

A developing topic of research in natural language processing (NLP) is sentiment analysis (SA) It's also known as "opinion mining." Automatically detecting attitude or sentiment in a text is often helpful. SA can determine whether a statement is positive, negative, or neutral. As a result, SA has recently attracted a lot of interest in fields including politics, business analytics, customer reviews, and other academic disciplines. Sentiment analysis is the automation of the process of identifying and categorizing the reviewer's expressed sentiments in a written text on a particular subject or item. It enables mining of the massive amounts of data shared on social networks. People express their thoughts on social media, review websites, and blogs; as a result, SA's mission is to analyze the provided textual data in order to identify the emotions inside. As an example, on an online store, customers generally leave feedback after buying a particular item. These reviews are typically extremely helpful for other buyers who want to purchase the product. However, because there are so many evaluations, it is impossible for customers to read them all and make the best choice. Sentiment analysis makes it possible to automatically analyze this massive amount of material and draw out people's opinions. Emotional analysis enables a deeper understanding of the individual emotions behind those basic emotions that are often tied to behavior by categorizing the data into whether it shows a positive or negative feeling. Identification of potentially harmful societal attitudes and behaviors requires such understanding. For instance, a large number of remarks expressing the emotion "anger" may result in violent acts in the neighborhood. Sentiment analysis was done at both the sentence and document levels by the researchers. The SA at the document level has been used to categorize the attitudes expressed in the document, whether they are positive or negative. The models used to identify the attitudes expressed just in the studied sentence at the sentence level. Although the majority of sentiment analysis research has focused on English, more recent initiatives have turned their attention to other languages, such Arabic. To get the required results, many researchers combined machine learning techniques. However, before using any machine learning algorithms on these models, a lot of feature extraction is needed. Experts have always employed opinion mining approaches with machine learning to achieve acceptable results, but as neural networks and deep learning have gained popularity, experts are switching to this strategy to achieve superior outcomes.

Machine Learning (ML) is an artificial inelegance method that allows computers to learn from previous knowledge. Machine learning procedures use computational approaches to learn useful information directly from unstructured data without relying on a predetermined calculation as an approach. A sophisticated subfield of machine learning is deep learning. One of the best machine learning techniques for NLP problems is deep learning. It learns, forecasts, and suggests the ideal course of action for an application using composite algorithms. Deep learning can occasionally perform with accuracy comparable to the human mind or even surpass human intellect.

1.2 Problem definition

Sentiment analysis is one of the important areas that is rapidly expanding thanks to the growing online market. The goal of sentiment analysis, also known as opinion mining, is to identify the reviewer's feelings toward a topic using textual context. The problem statement can be described as the finding reviews polarity to be classified as negative, positive, or neutral. The fact that Arabic has been studied much less than English serves as a wake-up call for all Arabic researchers. However, in the last few years we can see a clear effort in Arabic sentiment analysis resources. Arabic has more challenges than western languages due to its morphology and the existence of many dialects.

This research further investigates how important it is to benefit from deep learning for discovering people's opinion on a large scale such as in social media. The research issue is evaluated by comparing the performance of different deep learning classifiers with a number of three datasets, the results will help us understand how effective deep learning models are in analyzing Arabic text and calcifying it is positive, negative, and neutral.

The paper attempts to answer the following questions:

RQ1 How dose different deep learning models perform on different dialect Arabic datasets?

RQ2 What is the best performing tested model for SA?

RQ3 What is the effect of preprocessing on the quality of Arabic sentiment analysis?

1.3 Thesis Contribution

This dissertation objectives are to fill the research gap for Arabic sentiment analysis and focusing on the dialect Arabic. It presents and test different deep learning calcification models for social media sentiment reviews. We also look at how to improve these DL models to get better results for the SA. We provide a new dataset for Arabic sentiment analysis for movie reviews.

1.4 Dissertation Organization

The dissertation is organized as follows:

- **Chapter 1: introduction:** gives an overview of the sentiment analysis
- **Chapter 2: Related work:** presents a summary of the research that has been done in this area.
- **Chapter 3: Literature review:** review of sentiment analysis and the fundamental concepts involved. It examines the many sentiment analysis technique designs, the importance of sentiment analysis, and its challenges.
- **Chapter 4: Dataset collection:** illustrates the datasets that are used in this research.
- **Chapter 5: Methodology:** describes the methodology for data collection.
- **Chapter 6: evaluation and result:** Displays the experimental findings that we came at by contrasting our suggested strategy with pertinent, current methodologies.
- **Chapter 7: conclusion and future Work:** it concludes the dissertation and provide potential directions for further research.

CHAPTER 2 RELATED WORK

In order to represent the grammar as a collection of words, Bashir et al. (2018) employed CNN utilizing a single layer of convolution. CNN is used in projects including image analysis, handwriting recognition, audio recognition, and text processing. Similar to this, Abu Farha & Magdy (2019) proposed a hybrid model in which CNNs were employed for feature extraction while LSTMs were used for sequence and context interpretation. Modern research on a range of Arabic dialect datasets, such as SemEval 2017 and ASTD, is produced under the current "Mazajak" technique for Arabic SA.

In addition, AlKhatib et al., (2020) Using CNN and the China-US trade war as a case study, a sentiment analysis of 52 economic opinion leaders was conducted. Every convolution has a filtering matrix, hence various convolutional operations of varying sizes were performed. Because each CNN layer is coupled to a max-pooling layer, which serves the purpose of using the significant (n-grams) attributes individually, this system performs well with SA tasks. The CNN experiment employed four sentiment categories. According to the results, CNN's sentiment analysis yielded classification results with an accuracy rate of 86%

Ibrahim et al. (2019) also analyzed reviews using a deep semantic analyzer built on an RNN. They suggested a movie recommender that analyzes reviews' emotional content in order to provide a better listing while taking the customer's preferences into account. They employed the method of categorizing movie reviews and assigning semantic emotions to them. Similar research was conducted by Rehman et al., (2019), who used LSTM and CNN models on a variety of NLP tasks and produced excellent results. In addition to max-pooling layers, the CNN model also employs convolutional layers to successfully retrieve higher-level information. However, the LSTM function can look at long-term connections between word sequences. The researchers used a hybrid model called the Hybrid CNN-LSTM Model, which merges LSTM and a very deep CNN approach, to solve sentiment analysis issues. The recommended Hybrid CNN-LSTM model outperformed the conventional deep learning and machine learning techniques in terms of precision, recall, f-measure, and accuracy. Along with the Amazon movie reviews corpus, they also employed the IMDB movie reviews corpus.

Two strong Arabic deep BERT models, ARBERT and MARBERT, were introduced by Abdul-Mageed et al. in 2020. To facilitate learning transfer on MSA as well as Arabic dialects, ARBERT and MARBERT were previously trained on enormous and varied Arabic datasets. ARBERT was pre-trained on 61GB of MSA text as a result (6:5B tokens). While MARBERT is a thorough pre-trained ML model focused on both MSA and Dialectal Arabic (DA), MARBERT was trained using sample 1B, a large-scale randomly chosen tweet.

Al-Twairesh et al. (2017) outline the steps taken to compile and generate a sizable collection of Arabic tweets in their work. To prepare and clean the gathered dataset is explained. From this dataset, a corpus of Arabic tweets with sentiment annotations was taken. Tweets in MSA and the Saudi dialect make up the great majority of the corpus. The corpus was manually annotated for sentiment. The annotators attended an hour-long training session and received annotation guidelines.

CHAPTER 3 LITERATURE REVIEW

3.1 Background information

Sentiment analysis (SA) is a growing research area in Natural Language Processing (NLP). It is also referred to as opinion mining. Generally, it helps to automatically determine attitude or emotion in a text. SA can identify a positive, negative or neutral statement. Hence, SA gained a lot of attention recently in areas such as politics, business analytics, consumer reviews, and other areas of studies. According to Aydln & Güngör, (2021) Sentiment analysis is the automation of the task of recognising and classifying sentiments that are stated by the reviewer in a written text about a certain topic or an object. It allows to mine the huge amounts of information shared on social networks. People express their opinions in social media, review sites and blogs, therefore SA is the task of analysing the given textual data to extract the emotions in it(Bodapati, Veeranjaneyulu & Shaik 2019a).To illustrate in an online shopping site users typically write their opinion after purchasing one of the products. Typically, these reviews are considerably beneficial for other customers who want to buy the item. Despite the huge number of reviews, it is not possible from customer to read all reviews to make the correct decision. Sentiment analysis allows to automate the process of analysing this huge amount of text and extract people's opinion from it. Emotional study enables for a deeper comprehension of the particular emotions underneath those overarching sensations by categorizing the data into whether it exhibits a positive or negative feeling such as, emotions that are frequently associated with personality (AlKhatib et al. 2020). Identification of potentially harmful societal attitudes and actions requires such understanding. For instance, a large number of remarks expressing the emotion "anger" may result in violent acts in the neighborhood. Researchers performed Sentiment analysis on both the document and sentence level. The document-level SA has been performed for classifying the sentiments stated in the document whether its positive or negative. while, in the sentence level, the models used to identify the sentiments conveyed only in the analyzed sentence (Dashtipour et al. 2021a).

Although the majority of sentiment analysis research has focused on English, more recent initiatives have turned their attention to other languages, such Arabic. Many researchers used a combination of machine learning techniques to achieve the desired results.

However, these models require lots of feature extraction before applying any machine learning techniques (Dashtipour et al. 2021a). Experts have always employed opinion mining approaches with machine learning to achieve acceptable results, but as neural networks and deep learning have gained popularity, experts are switching to this strategy to achieve superior outcomes. (Barrón Estrada et al. 2020).

3.1.1 Types of Arabic languages

According to (Rieser & Refaee 2014)There are three major variants used for Arabic language:

1. Classical Arabic (CA) This is the traditional Arabic. The Holy Quran is written in this language. Arabic literature, formal media, and Islamic texts are now commonly used. Typically, this variation is taught at Arabian institutions, including schools and universities.

2. Modern standard Arabic (MSA) A direct descendent of CA is. Today, it is the language most frequently used in discourse and communication, contemporary literature, and popular media including TV, radio, newspapers, and the Internet. Since MSA is a condensed version of CA, it is widely accepted throughout the Arabic-speaking world and is understandable to the majority of Arabians.

3. Dialect Arabic language (DA), it is a regional based dialect and is used in everyday communications. This includes informal communications, regular speech, and culture media. There are numerous Arabic dialect groups which can be sorted according to the region, which include gulf area, Egypt, levant and Maghrib.

3.2 Techniques and Applications for Sentiment Analysis Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis (also called feature-based sentiment analysis) is the study that concentrates on recognizing all sentiment terminologies inside a given data and the aspects to which they state. In many cases people have different opinion about each aspect, these entities will have many aspects or attributes(Kalpana B et al. 2022). This case can be visible when analysing reviews written about a product as cameras, smartphones or and other product in the market. In fact, it can be in any case where people have different opinion such as in discussion forms. People sometimes will mention the pros and cons of an aspect. Therefore, classifying the review as either positive or negative would result in missing the valuable information summarised in it.

According to (Kalpana B et al. 2022) Sentiment evaluation involves reviewing online conversations such as tweets, written posts, or comments about some offers or topics and separating the sentiment of a person. Depending on the purpose of the evaluation sentiments can labelled accordingly such as emotion detection, issue-primarily based sentiment evaluation, and purpose evaluation. Aspect-based sentiment analysis (ABSA) is a text-based content evaluation method that classifies statistics via issue and finds the sentiment attributed. ABSA is a sentiment evaluation that enables the growth of the marketing in an organization by understanding the abilities of their product which they want to improve using with customer’s remarks.

3.3 Machine Learning Approaches

Machine Learning (ML) is an artificial inelegance method that allows computers to learn from previous knowledge. Machine learning procedures use computational approaches to learn useful information directly from unstructured data without relying on a predetermined calculation as an approach. The ML regularly improves their performance depending on the number of training data obtainable for learning process. Deep learning is a specialized form of machine learning. There are two types of ML approaches: supervised learning, and unsupervised learning. Supervised learning uses algorithms to train a model using identified input and output therefore it can predict upcoming outputs. unsupervised learning, finds useful patterns and basic structures in input information. The popular ML classifiers used by researchers were Support Vector Machines (SVM), Naive Bayes (NB), Ridge Regression (RR), Random Forest Trees, Maximum Entropy (ME), and Adaptive Boosting (AdaBoost) (Hasan et al. 2018). the accuracy of sentiment evaluation and estimations can be attained by behavior analysis built on social networks.

According to a test done by Alkhatib et al., (2019) they applied a number of classification models which shows that Polynomial Networks and SVM achieved the best result regarding Arabic text classification, reaching classification accuracy of 96.49% and 94.58%. The results also presented that the use of stemming resulted in drawback in terms of response time, and that the larger the dataset used, the better the classification accuracy will be. The researchers applied two metrics for the evaluation the response time and the F1 Score. The response time represents the period consumed by the algorithm for applying classification model. This metric emphasizes the standing and efficiency of the model in real-time applications. knowing the model accuracy is significant since it allows us to evaluate how accurately a given model will predict test data.

3.3.1 Challenges

For example, the word "loud" might be used negatively in a review of a washing machine yet positively in one of a headset. The polarities of certain expressions can, however, differ between languages. Therefore, lexicon-based or domain-based techniques are required to address these difficulties. (Aydln & Güngör, 2021).

Another challenge is the hyperbole, it involves expressing ideas and sentiments with an overstated meaning; In Arabic it is called صيغة مبالغة (Baly et al., 2017). It is frequently employed to make a forceful statement or to stress a meaning. Euphemism which in Arabic الكناية reduces the specifics of a statement or viewpoint that seems harsh in order to clarify the underlying sentiment's meaning.

Morphology is a knowledge that studies originating branches of a word referring to their origins; the requirements of the structure of a word in relations to abstraction and increase (Aydln & Güngör, 2021).

3.3.2 lexicon-based techniques

Lexicon based methods or corpus-based methods control the vocabulary and word meanings in a text. Many researchers applied a hybrid sentiment analysis approach which uses a combination of lexicon-based and corpus-based techniques. The objective is to classify the input sentiments that are not classified using the lexicon-based approach. Assiri et al., (2018) proposed a new lexicon-based approach for Saudi dialect sentiment analysis that uses domain independence. They developed a new annotated dataset for Saudi dialect to be used with large scale lexicon. In many experimentations, The experiment's built-in corpus is evaluated using ML algorithms. In this research, by applying negation rules the accuracy reached 76.5%. the results also shows that there is a solid relation between polarity text and non-polarity text in the Saudi dialect and other Arabic dialects. For example,(Almuqren & Cristea 2021) built dataset for Saudi dialectal Arabic SA, and used a number ML algorithm to test the performance of the dataset

3.3.3 Supervised Learning

In the face of uncertainty, supervised ML develops a model that forecasts future results based on evidence. A supervised learning model trains a model to generate accurate estimates for the incoming data using a predetermined set of input data and known responses to the output. The methods that rely on labeled data are those based on supervised learning in machine learning. There are two types of supervised learning techniques, classification and regression techniques. Examples of classification models are Logistic Regression, SVM, NN and Bayes. The majority of studies concentrate on supervised learning techniques that extract sentiments based on reviews. Saroufim, Almatarky, and Abdel Hady (2018) create language-independent sentiment-specific embeddings. On base of a word2vec model, they use a supervised component. Additionally, the majority of research focuses on supervised learning frameworks and generates sentiments on a review basis when using unsupervised and semi-supervised approaches on text sources. The supervised methods typically apply boolean or TF-IDF metrics as a weighting technique(Aydln & Güngör 2021)

Classification techniques predict separate feedback; for instance, it can assume whether an e-mail is authentic or spam. Or in medical situation such as, if a tumor is cancerous or benign. Classification approaches classify input information into groups. For instance, classification techniques can be used in handwriting recognition where it uses use classification to identify characters and numbers. Regression techniques predict constant responses; for example, measurable physical numbers such as power or battery condition. It can also be used for mustering life time of piece of equipment and predict expected failures.

In Arabic sentiment analysis the performance of the module is mainly measured by prediction accuracy. Therefore, researchers preformed several ML algorithms and others used a combination of different methods. To perform ASA, numerous machine learning algorithms have been trained. The accuracy and the predictions is typically used to gauge how well they function. According to a study done by Duwairi and El-Orfali (2014) the top three classifiers that regularly displayed superior result were SVM, k-nearest neighbor (KNN), and NB. Researchers also discovered that the pre-processing step improved the classifiers accuracy for Arabic sentiment, opposed to un processed sentiments. Many researchers targeted only the modern standard Arabic (MSA) however, researchers now are targeting dialectical Arabic considering its widespread and presence in social platforms (Oueslati et al. 2020). ML involve adding classifiers, utilizing data features including unigrams or bigrams, where it can include part-of-speech tags, and are fundamentally a supervised classification module (Zhang et al. 2019).

Some of the most used ML techniques in sentiment analysis are Support Vector Machine (SVM), Regression, Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and K-Nearest Neighbor (KNN). Besides, the most commonly used supervised approach is the support vector machines (SVM), where it is able to avoid overfitting. This can happen using the kernel trick and being identified by a convex optimization problem where the matter of local minima does not exist. Support Vector Machine (SVM) is associated to the supervised learning method and meant for classification and regression techniques.

In research done by Barrón Estrada et al., (2020). They created a cuprous for educational purpose in the field of programming language. They compared different sentiment analysis techniques which are ML, deep learning, and an evolutionary method called EvoMSA. The corpora expressed the emotions of students regarding instructors, examinations, assignment, and academic projects. For ML classifiers we decided used the Bernoulli and Multinomial Naïve Bayes methods, k-nearest neighbors (KNN), Support Vector Machine and Linear Vector Machine. The tested ML model scored a respectable and accurate 92%. However, (EvoMSA) produced the greatest results, with an accuracy of 93%.

A. Al Shamsi & Abdallah (2022)employed the several classifiers such as logistic regression (LR), multinomial Naïve Bayes (MNB), (SVM), decision tree (DT), random forest (RF), multilayer perceptron (MLP), AdaBoost, GBoost, and an ensemble model. the dataset used contained a number of 70,000 comments obtained from Instagram comments, where most of the comments were written in the Emirati dialect. The applied the classifiers after converting the dataset into Tfidf Vectorizer. They used python library including SciKit-Learn and NLTK Tool Kit. These libraries allow programmers to have access to many machine learning techniques and free and with many open-source models. Four important measures were used to assess performance the accuracy, Recall, Precision, and F1-score. The results show that the combined model, such as combination of RF, MNB, LR, SMV, and MLP algorithms, performed better than other ML models which reached accuracy of 80.80%. Additionally, the RF model achieved the highest result in metrics of recall and F-measure, whereas, the MNB achieved the highest result in metrics of precision

3.3.4 Unsupervised Learning

Unsupervised learning identifies fundamental data structures or hidden patterns in the data (Aydln & Güngör 2021). There are several research obtain the emotions in a sentence or document based on the sentiments of words using unsupervised method(Aydln & Güngör 2021). It is employed to infer conclusions from data having input data and answers without labels. One of the most known unsupervised learning techniques is clustering. Therefore, Clustering technique is used in data analysis to discover hidden patterns or groups in data. While ML-based methods are supervised methods that depend on labelled data, Lexicon-based methods use unsupervised techniques that do not. The number of words and their meanings in a text are controlled by lexicon-based approaches or corpus-based techniques.

Aydln & Güngör, (2021) applied binary sentiment categorization model on Turkish text on a document level. They made a complete analysis and recommended a framework that involved unsupervised, semi-supervised, and supervised methods, as well as the combination of these methods. The methods were applied on Turkish twitter dataset for movie reviews. The tweets were annotated manually by two professionals and the data was compiled. In addition, they evaluated their approach on an English dataset which produced greater results. The researchers claim that their method can be applied on other languages as well. Semi-supervised methods reserves that join sentiment data were often used in SA experiments.

(Aydln & Güngör 2021) concluded that combining unsupervised/semi-supervised and supervised methods produces the top outcomes in both datasets. This demonstrates the need to include knowledge obtained through an unsupervised method into the classification process in order to obtain more useful findings

3.4 Deep learning (DL)

Deep learning is an advanced sub failed of machine learning. Deep learning is one of the most effective ML methods for NLP tasks. It uses composite algorithms to learn, predict, and suggest the best possible outcome for an application. DL sometimes preformed as accurate as human mind or it can also exceed human intelligence. It can be accurate as the human brain, or sometimes it exceeds the level of understanding of humans. Deep learning consists of input layer and an output layer in addition to hidden layers in between where, each hidden layer uses differ algorithm (Sun, Luo & Chen 2017). The most common deep learning techniques are CNN, RNN, DNN, LSTM, and GRU. Neural networks achieved popularity and succuss due to it surpassing traditional ML models such as KNNs and SVM (Ibrahim et al. 2019). deep learning algorithm also includes in multiplication and weight selection steps.

According to Bodapati et al., (2019) DL approaches outperformed the popular ML methods such as, logistic regression, SVM , MLP . (Oussous et al. 2020a) proposed framework for improving Arabic sentiment. The provided an Arabic corpus called Moroccan Sentiment Analysis Corpus (MSAC). The results show that using deep learning methods outperforms the traditional machine learning approaches such as SVM, naive Bayes classifiers and maximum entropy. Considering the challenges in Arabic language the researchers aimed to improve Arabic sentiment analysis accuracy. The dialect used in this research is the Moroccan and general Arabic. The Corpus (MSAC) is characterized by informal language such as repetitive letters and non-standard shorts. The goal of this work is to create an ASA structure that enhances the SA accuracy of Arabic language. The MSAC has 2000 annotated reviews where 1000 are positive and 1000 are negative.

As stated in Aydln & Güngör, (2021) Between the supervised methods, SVM are generally preferred because they can prevent overfitting by using a kernel method and are defined using a convex optimization problem, which eliminates the issue of local minima. It is also claimed that DNNs have recently become increasingly well-liked in a number of fields because to their precision, effectiveness, and adaptability. While most classical ML algorithms take a lot of time and effort to extract characteristics. However, these traits can be automatically retrieved using DNNs. It has been used in word2vec sentiment analysis projects that use word embeddings.

3.4.1 Convolutional Neural Network (CNN)

A convolutional neural network (CNN) is one of the powerful methods for deep learning. CNN's analysis the data using several layers which, eliminates the need for human extraction. CNNs are a powerful tool for NLP which showed superior results (Alayba et al. 2018a). CNNs are commonly used in computer vision tasks. It can be performed on multiple images, texts, audios, and videos. According to Ibrahim et al., (2019). The CNN is widely used for pattern identification in hierarchal data classification.

CNNs are a mathematical function that are stimulates human neural networks. Some just utilize the two primary layers, input and output layers, but others can employ three simple levels, such as input, output, and hidden layer. (Ibrahim et al. 2019). There has been a lot of research in the last few years on NLP and sentiment analysis which shows good results (Alayba et al., 2018). According to Oussous et al., (2020), it is clearly stated that the CNN and LSTM algorithms outperformed the single model (SVM, NB and ME). Bashir et al. (2018) applied CNN using a single layer of convolution the purpose of this technique is to represent the syntax as a collection of words. Implementations of CNN can be seen in works such as analyzing images, handwriting recognition, speech recognition, and text calcification. CNNs are widely used in analyzing images to find features such as patterns, objects, time series, faces and analyze the whole scene. Useful information can be retrieved from images which sometime cannot be discovered by humans. A CNNs uses many layers that each learn to detect diverse features from the provide data. According to El Bazi & Laachfoubi (2019) CNN are considered a type of neural networks where it uses convolution in place of general matrix multiplication in one or more layers. The convolutional layer is the fundamental layer of a CNN, and it is where most of the computation takes place. It involves some components, such as input data, a filter, and a feature map. The convolutional layer can be embodied as a layer of a neural network as each neuron can represent any function. Among any two matrices a grouping function occurs using convolution layer. In the CNN function the Convolutional layer is mostly the first layer. El Bazi & Laachfoubi (2019) applied CNNS where to compute the character representation for words. In Gehrmann et al. (2018) experiment of deep learning, the results show that CNNs are an effective substitute for current methods in patient phenotyping and cohort identification.

AlKhatib et al., (2020) 52 economic opinion leaders were subjected to a sentiment analysis utilizing CNN, with a case study of the China-US trade conflict. Every convolution contains a filtering matrix, therefore, different size of convolutional operations was executed. This system works best with sentiment analysis tasks because each layer in the CNN is connected to a max-pooling layer where its function is to use the most important (n-grams) characteristics separately. Four sentiment categories support, strong support, dissent, and strong dissent were used by the CNN classifier in the experiment. The findings indicate that the classification accuracy produced by the CNN employed for sentiment analysis was 86%. Support Vector Machine, or SMV, is another classifier. Five emotions—angry, depressed, thrilled, pleased, and worried—were examined by the clarifier. An accuracy of 82% was attained by the SVM model used to perform deep emotional prediction while distinguishing between five feelings: anger, depression, excitement, happiness, and worry.

3.4.2 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) is another popular deep learning approach which is commonly used in NLP. RNN is broadly used with applications that predicts text or time. There are several reasons that makes RNNs one the most used deep learning for sentiment analysis and NLP generally. RNNs are implemented effectively with sequential data such as text. Additionally, they can be used to recurrently predict sentiments in a text as each token in a part of text is addressed. After the model reads all tokens; This way the model is trained and can predict future sentiments. RNNs can also be enhanced by the integration of an attention mechanism, which is a distinctly trained component of the model. this benefits the model to control on which tokens in a sequential script to apply its focus. Similar to CNNs it can RNNs be used effectively applied on printed text and achieving superior results. In contrast, the algorithms used in RRNs can understand the connection between words in a text where it can study the words before and after and not just a single word (Elnagar, Al-Debsi & Einea 2020). In contrast to CNNs, RNNs are constructed using many layers (input, hidden, and output layers) (input, hidden, and pooling layers). The CNN is frequently used in hierarchical data classification to identify patterns. While RNNs are more frequently utilized with written input in NLP for semantic analysis and classification (Ibrahim et al. 2019).

RNN uses algorithms that will find a relationship between words in a context; however, it is different from other neural networks. It will not only look at a single word; it will study the words before and after (Elnagar, Al-Debsi & Einea 2020). It can function using bidirectional style with different length of text; additionally, it focuses on both previous and following elements. as stated by Al-Ayyoub et al. (2018), bidirectional RNNs is a function where is contains of dual RNNs loaded on top of each other, and the outcome depends on the hidden layer of the two RNNs. According to Bodapati et al., (2019) RNN generally provide state of the art performance in application such as, machine translation, caption generation and language modelling. However, they suffer from the vanishing or exploding gradients issue when used with long syntax. According to Ibrahim et al. (2019) utilizing back propagation through time, RNNs are trained (BPTT). As a result, in extended sequences, the gradients may explode or disappear when it flows back through time, making it impossible for the algorithm to learn. This is the reason why RNNS are not accurate when modelling long sequences in data. They also used a deep semantic analyzer built on the RNN to analyze reviews. They proposed a movie recommender that analyses emotions on movie reviews to recommend a better listing considering the preference of the client on a targeted mobile application. They used the technique of grouping of movie reviews and labelling them with semantic emotions.

RNNs have been used for this SA applications to make the most recent progress on this topic. RNNs are a special kind of neural network that can simulate sequence data. Although RNNs are capable of providing cutting-edge performance for tasks like SA, language modeling, caption generation, and machine translation. The outcome of the model is determined by vector of each token repetitively. RNNs suffer from some limitations including vanishing gradient. Therefore, researches developed upgraded versions of RNNs that are assigned to achieve their requirement.

3.4.3 Long-short Term Memory (LSTM)

Long-short term memory (LSTM) networks are extended versions of RNN networks. It is used mainly to solve the issue of vanishing gradient in RNN. LSTMs are best in working with very long sequence data. The LSTM uses additional memory-cell to the model. It is intended to solve some problems of RNNs, such as in the vanishing gradient case. LSTM function differently to calculate the hidden states despite having similar architecture to RNNS. In particular, the variance is that LSTM can recall data from farther previous layers, such as learning the sentences in different paragraphs (Elnagar, Al-Debsi & Einea 2020). Also, long-distance dependencies can be easily captured using LSTM. The LSTM determines whether to discard or forward an information using the functions forget and update gates. In cases where there is more information needed, LTSM can outperform the traditional RNN.

GRU (gated recurrent unit) functions similar to LSTM model however, the quantity of gates is decreased in GRU by integrating the input data and forget gates (Bodapati, Veeranjaneyulu & Shaik 2019). GRU has in faster training period compared to LSTMs due to having fewer number of parameters

In a work done by Bodapati et al., (2019). they use LSTMs to determine the SA for the analysis of a movie review task. The research was centered on predicting the polarity of selected movie feedbacks by classifying it positive or negative. They look at how hyperparameters such as, dropout, layer count, and activation functions affect the system. They used variant of neural network configurations and noted each model's performance depending on its configuration. The reported that LSTM model had the best performance when used on the selected dataset which is the IMDB datasets. LSTM was compared with other methods which are logistic regression, SVM, MLP and CNN. They also stated that SVM is the best shallow classification model compared to the other shallow models. The dataset used holds a total of 50,000 reviews where, 25000 are labelled positive, and 25000 are labelled negative.

A similar work done by Rehman et al., (2019). where they also used LSTM and CNN models on several NLP tasks which achieved outstanding results. By using convolutional layers in addition to max-pooling layers, the CNN model successfully recovers higher-level information. The LSTM function, however, may examine long-term relationships between word sequences. To overcome sentiment analysis problems, the researchers utilized a hybrid model termed the Hybrid CNN-LSTM Model, which combines LSTM and a very deep CNN technique. In terms of precision, recall, f-measure, and accuracy, the suggested Hybrid CNN-LSTM model demonstrated superior performance to the traditional deep learning and machine learning techniques. They similarly used the IMDB movie review corpus in addition to Amazon movie reviews corpus. In a work done for Parisian language, Dashtipour et al., (2021). proposed a new, context-aware, deep-learning-driven method. The suggested manual-feature-engineering-driven, SVM-based technique was compared to CNN and LSTM. According to experimental findings, the LSTM algorithm outperformed the MLP, autoencoder, SVM, logistic regression, and CNN algorithms in terms of performance. The experiment also anticipated deep learning concludes the significance of single layers in the perspective of the whole approach. Specifically, the bidirectional-LSTM reached the highest accuracy of 95.61%, using the movie corpus. However, the 2D-CNN outperformed LSTM when used with the hotel corpus.

Conforming to Hasan et al. (2018), emojis were used to extract feelings from text through distant supervision and a bidirectional -LSTM. This technique was used to execute multiclass sentiment classification. It also spots scam text in reviews via a slightly improved version of this neural network. They used a two-layer bi-LSTM framework with attention over the last layer for the message-level classification application. Regarding the topic-based classification section, bidirectional LSTM is improved using a context-aware attention tool.

3.4.4 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a multi-layer bidirectional transformer encoder that was released Devlin et al. (2019). Transformers were introduced to address the problem of the long-range dependency challenge. Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model. It is a deep learning structure to be used for the tasks of NLP. BERT is also considered a neural-network-based approach for NLP pre-training. It is funded on Transformer and its core principle - attention, therefore it recognizes the context relationship among different words. Moreover, it can be utilized to distinguish the connection of words in a context or queries. Unlike previous language representation (LR) models

BERT is pre-trained by reading both left and right context for all layers (Chouikhi, Chniter & Jarray 2021). The main component of transformer are a set of encoders and decoders. Encoders encode text into a representation which then is decoded by the decoders. The output of the last encoder will be the input of initial decoder. The decoder then preforms certain transformation tasks and then send to the following decoder. BERT model only encodes information and creates a language model, so decoding not required. Similar to vectorization BERT converts text into numbers. This phase is an important stage in all machine learning tasks. This is because ML models deals with text in numbers or vectors format as an input to the model. It is a vital task for training the model. Therefore, BERT models are meant to transform all text data to be recognized with other forms of data. Such as in the task of making predictions in sentiment analysis. Unlike RNN, CNN, or any directional models that read text sequentially, BERT models analyze the whole text and surrounding words to understand the context. Usually, BERT models are trained using a huge volume of datasets to learn words relationships, which then allows it to have an advantage over other ML techniques. BERT is a deep bidirectional, unsupervised NLP which is trained using a large textual dataset. Applying BERT toan NLP task requires only to modify one added output layer for the downstream task(Chouikhi, Chniter & Jarray 2021). The most common training approaches in BERT are **Masked Language Modelling (MLM)** and **Next Sentence Prediction (NSP).** The main target is to overcome the dependency challenge. According to many researcher BERT models achieved state-of-the-art outcomes when used for NLP tasks.

By learning to anticipate text that might occur before and after (bi-directional) other text, Google developed this algorithm in 2018 to enhance contextual understanding of unlabeled text across a wide range of activities. The ability for multilingual BERT to operate in a variety of languages is its key advantage. However, it was proven that Arabic specified BERT are more efficient since it is using a large size of corpora set.

Abdul-Mageed et al., (2020). introduced two powerful Arabic deep bidirectional transformer-based models ARBERT and MARBERT. ARBERT and MARBERT are pre-trained on huge and diverse Arabic datasets to enable learning transfer on MSA along with Arabic dialects. Since Arabic multiple of varied dialects, majority of such dialects are under-studied due to limitations of resources. Multilingual models like mBERT and XLM-R were pr-trained generally on MSA text, which is similarly the case for AraBERT and ARBERT (Abdul-Mageed et al., 2020). Therefore ARBERT was pre-trained on 61GB of MSA text (6:5B tokens). Whereas MARBERT is an extensive pre-trained ML model centered on both

Dialectal Arabic (DA) and MSA. MARBERT, was trained using a large scale randomly selected tweet, which sample 1B tweet. MARBERT use the similar BERT base network architecture as ARBERT. However, since tweets are short, the model did not require to use next sentence prediction (NSP). Arabic BERT is a group of BERT language models made up of four models of various sizes that were developed employing entire word masking and masked language modeling (Safaya, Abdullatif & Yuret 2020). Models with large, base, medium, and mini sizes were trained with the same data for 4M steps. Among the newly made ULM models, Devlin et al., (2019) built a multilingual language version BERT employing 104 languages including Arabic nevertheless, this model has only been evaluated on Arabic sentence contradiction problem.The main benefit for applying multi languge BERT is that it can be running on various languages. Yet, one significant limitation for multi-lingual BERT is that it was restricted to running alongside multiple-language corpora. This means that it did not benefit from the far bigger datasets that were available for Arabic, which reduced its inherent representation for Arabic.

Another Arabic BERT is the hULMonA model by Eljundi et al., (2019) which has also produced state-of-the-art outcomes. Its primary objective is to enhance Arabic NLP task performance and generalization capabilities by creating new universal language models (ULMs) for Arabic. hULMonA is the first Arabic specific ULM model, Moreover, the also demonstrate how multilingual pre-trained BERT may be adjusted and used for Arabic classification tasks such as SA.

In an effort to matching the success that BERT had with the English language, Antoun et al., (2020) pre-trained BERT particularly for the Arabic language. AraBERT's performance is contrasted with that of Google's multilingual BERT and other state-of-art methods. The outcomes demonstrated that the recently created AraBERT performed at the high end on the majority of Arabic NLP tasks. Their pretrained araBERT models are freely offered on GitHub to inspire Arabic NLP research. Both MSA and DA were present in the datasets that we considered for the subsequent tasks. As a result, AraBERT outperformed multilingual BERT. Despite having been trained on MSA, AraBERT performed well on dialects that had never been encountered previously. SA, named entity recognition, and question answering were the three downstream tasks we used to assess ARABERT's comprehension of Arabic. As a starting point, we contrasted ARABERT with the multilingual edition of BERT as well as other cutting-edge outcomes for each work.

By making the model anticipate the entire word rather than just part of it, whole word masking enhances the pre-training work. We also use the Next Sentence Prediction (NSP) task, which can be helpful for many languages comprehension tasks like Question Answering because it teaches the model how two phrases relate to one another.

According to Chouikhi et al., (2021) also proposed a new method for Arabic BERT where they By replacing the standard BERT Tokenizer With an Arabic BERT Tokenizer. Testing was conducted using various types of Arabic such as DL, and MSA. To get the best outcome with various datasets, they employed casual search with hyperparameter optimization approach. Comparing the proposed technique to existing Arabic BERT and AraBERT models, the experimental study demonstrates the effectiveness of the suggested method in means of classification quality and accuracy. Moreover, in a study done by Safaya et al., (2020) they introduced a hybrid BERT-CNN model. This study examined how well BERT-CNN could detect objectionable speech in social media content and compared its structure to that of other models. It has been demonstrated that employing BERT with CNN produces improved outcomes compared to using BERT alone, or CNN alone. Furthermore, the ArabicBERT pre-training procedure was described. Using least amount of text pre-processing the model was able to accomplish impressive results. Therefore, in the OffensEval2020 which is a part of the SemEval 2020, the team was ranked among the top four teams for all languages included. The achieved a macro averaged F1-Score of 0.897 in Arabic. The ArabicBERT model is available online for free for the purpose of research.

In the research done the results demonstrate that the transformer-based BERT model outperformed the traditional model in terms of emoji prediction accuracy. As a result, our studies with a complex design show that our strategy increased the score above the standard approach in dual languages. The BiLSTM model with focus displayed a strong classification performance. Our findings, however, demonstrated that the implementation of the Attention BiLSTM architecture was outperformed by the BERT model, which is based on a transformer.

3.4.5 Hybrid model

A Hybrid model Is A method that combines different forms of DNN with prediction methods to model uncertainty. Diverse types of DNN, such as RNN or CNN, have produced outstanding results in terms of their efficiency and popularity for dealing with different ML applications and analyzing numerous forms of data. However, DL algorithms fall short of the probabilistic or Bayesian techniques in their ability to eliminate ambiguity. Hybrid learning algorithms therefore use two different models to reinforce the benefits of every. Bayesian deep learning, Bayesian GANs, and Bayesian conditional GANs are other instances of hybrid models.

By combining homogenous CNN classifiers, a hybrid DNN model is created. By adjusting the initialization of the neural network's weights and the input features' diversity, the ensemble of classifiers is created. A value of one is output by the convolutional neural network classifiers for the anticipated class and a value of zero for all other classes.

Abu Farha & Magdy, (2019) suggested a hybrid model in which LSTMs were used for sequence and context interpretation while CNNs were utilized for feature extraction. The current "Mazajak" approach for Arabic SA produces cutting-edge findings on a variety of Arabic dialect datasets, including SemEval 2017 and ASTD. Free access to Mazajak is provided for academic use.

3.5 Compare Deep Learning models

Compared to directional models such as RNN and LSTM which consider every input sequentially (left to right or right to left). Unlike previous language representation (LR) models BERT is pre-trained by reading both left and right context for all layers (Chouikhi, Chniter & Jarray 2021). In reality, Transformer and BERT are non-directional models. For instance, both of these models read the full context as an input opposed to sequential ordering. This feature grants the model to study the whole setting of a sentence while taking in concertation all other words in the sentence. Generally, transformer models rely on self-attention to compute data and represent input and output without relying on sequential ordering as RNNs or CNN. For transformer a Single encoder layer has a self-attention technique and a feed frontward neural network. Alternatively, a decoder has both the mechanisms of an encoder and an encoder - decoder attention. This concludes that BERT is a transformer-based model because it employs an encoder that is very comparable to the transformer's original encoder.

3.6 Pre-processing

The pre-processing is a significant phase in NLP and SA. It helps to improve the condition of the dataset collected and guarantees a better performance sentiment analysis. The common pre-processing steps include text cleaning, normalization, tokenization, stop words removal, and stemming. Depending on the study objective and the selected language, the prepressing step can be carried out in several stages. In a study done by Mhamed et al., (2021) they tested the performance of Arabic sentiment classification using novel pre-processing step. The procedures involve using a personalized stop list and handling emoticons. According to the findings, Arabic attitude categorization can be made more effective by using tailored stop lists and content terms in place of emoticons.

1- Data cleaning.

Data cleaning is a critical phase to improve data quality. It enhances the task of data polarity detection. After collecting the data, the cleaning step removes spelling errors and slang words. However, it is not easy to apply this in Arabic text since there are variation of dialects and type of Arabic language. According to Alayba et al., (2018). Arabic is a challenging language when dealing with linguistic domain for NLP. It has morphological difficulties and variety of dialect which demands advanced pre-processing. A useful process in cleaning social media corpus is to delete usernames, hash tags, URLs, punctuations and unwanted whitespaces. Depending on the task others remove long wors, special characters numbers and English text. Pre-processing and evaluation of Arabic unrefined contents is particularly important to decrease the text noise and enhance the efficiency of the dataset.

In this paper similar to Alayba et al., (2018) the pre-processing step was as followed:

1. Eliminating all English or none Arabic words.

2. Removing all digit numbers

3. Removing special characters while keeping the characters related to emojis since its describes and provides value the writers emotions. The characters removed were as (!@#$%^&) since it does not add value to the sentiment.

4. Emoticons

In social media users often use abbreviations, misspell word, use emojis and other symbols that express special feeling. Emoticons are interactive visual depictions of face expressions that typically use combination of symbols, characters, and images (Wolny 2016)

Emojis give people the ability to visually represent the equivalent of many emoticons. But certain emoticons don't have an equivalent in emojis, and vice versa in all areas (Wolny 2016). Additionally, at least on Twitter, the most popular emojis do not exclusively consist of face emojis.

2- **Stop words removal:**

Is the process of removing words that are used for structuring language while not contributing to the main context. As an example of these words are a, are, the, was. ext. These words are mostly frequent and does not improve the task of sentiment analysis or classification. This helps in reducing cuprous size without having to lose any of the main information.

3- **Tokenization:**

Tokenization is an essential step for most NLP tasks, where it divides the raw text string using whitespaces into a list of separate words. It separates a sentence or document into tokens which are words or phrases (Sun, Luo & Chen 2017). It is a simple task in languages the uses space to separate words sus as English, Arabic and French. However, it is more challenging in languages where wors are not separated such as Japanese, Chinese and Thai. tokenization can be used for Arabic language as most words are space delimited.

4- **Stemming**:

Stemming is an empirical step for removing word affixes according to some grammatical rules and converting them to an infinitive recognized form or stem. As an example, the word wait, waits, waiting and waited all become wait when stemmed. There are several of the most popular standard stemmer algorithms such as, Porter's stemmer and Snowball stemmer library2. The benefit of stemming is utilized to decrease the feature space.

Stemming is one of the most critical pre-processing steps in sentiment analysis besides being a popular task for natural language processing (NLP). As illustrated by Alayba et al., (2018). Stemming regulates words by converting each word to stem, base or the root form of a word. Further the stemming involves removing the affixes (such as infixes, prefixes and suffixes) from a phrase. Therefore, the application of stemming allows to decrease the corpus size into a small space (Oussous et al. 2020b). It will also improve the quality of text categorization; it is important to use the stemming technique. For instance, the root of the Arabic word teachers (معلمات) is (علم) and the stem is (معلم). As stated by Oussous et al., (2020) they used two stemming methods for Arabic data calcification, which are (Light10) and root-extraction methods (KHOJA).

5- Lemmatization:

A process known as lemmatization transforms a word into its dictionary-uninflected form. Although comparable to stemming, it is accomplished using a separate, more rigorous set of procedures that incorporate morphological analysis for each word.

6- **Stop words removal**:

Is by eliminating words that do not contribute to the word meaning or add any value to the context. Mostly these words are used for language structuring. As an example, are a, are, the, was, ext..

7- Normalization.

Another major pre-processing step for Arabic SA is normalization. It is the step of converting all the forms of a word into a common form. For example, the Arabic letters (وٌ ؤ وّ) is converted to the common letter و. Normalization benefits in ensuring a consistent form of the used Arabic text. (Oussous et al. 2020b) used PyArabic, which is an Arabic normalization library. The PyArabic allows basic functions to analyze Arabic text such as detection of Arabic letters, Arabic letters features, eliminating diacritics or tashkeel. In addition, it contains Tashaphyne library for word normalization where it finds for duplicated or repeated characters and replace them one character. The repetition in Arabic language expresses affirmation and accentuation

8- Vectorization and word embedding

The reviews written in social media or any other sources are written in text format which humans can easily understand. However, neural networks currently are unable to understand raw text format and cannot be used directly with the given text data. Pre-processing needs to be applied to convert the data to a format that neural networks can understand which is vector format. The text needs to be converted to a numerical format called vector. The data can be utilized as an input for the rewired model once it has been converted to vector form. Vectorization of the written material is the process of transforming text into a vector (Bodapati, Veeranjaneyulu & Shaik 2019). Nearby vectors ought to indicate words that are comparable to one another. Unstructured data can be subjected to vectorization, which uses mathematical calculations to represent every word in a text with a vector. Word embedding is the method of analyzing words and their meanings from a given text and represent the word as a vector form (find citation). The word vector is the projection of the word into a continuous feature vector space. Words that have similar meaning should be close together in the vector space as illustrated.

One of the mostly used word embedding techniques in NLP is Word2vec. The Word2vec approach is a neural network technology that generates correct representations without the need for labels by learning from word spread representations. The network creates word vectors with intriguing properties when given enough training data. (Djaballah, Boukhalfa & Boussaid 2019). Word2vec has two types, Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model. Basically, CBOW guesses the word depending to the given text, while Skip-gram guesses the text depending to the given word(Bodapati, Veeranjaneyulu & Shaik 2019a).

Rehman et al., (2019). trained primary word embeddings using the Word2Vec technique. According to the definitions of the terms, the Word2Vc converts the text into a vector of digital values, calculates the distance between words, and groups related words together. In addition to long-term dependencies, the model incorporates a collection of characteristics that were extracted using convolution and global max-pooling layers. The given model also practices dropout function, normalization and a corrected linear unit for accuracy enhancement. In another research, (Alayba et al., (2018).Built Word2Vec models using a corpus collected from different newspapers written in Arabic. On this corpus they applied machine learning functions and CNN using variation text feature extractions. The result showed improved accuracy on the task of sentiment classification. The best Word2Vec model was utilized to create an Auto Arabic Lexicon applied with different ML methods.

The Bag of Words (BoW) model is to represent the given text to a simpler numerical format called vector. The sentence is then represented as a string of numbers which the computer can understand. In addition, every component of the vector represents the presence or absence of a particular word. Also, frequency count is utilized to represent the number of occurrences for a particular word. As a replacement the model uses a binary 1 or a 0 for the representation. The result of BOW is a vectors size equal to the vocabulary size. Whereas in word embedding the vector size is considerably smaller than the size of the vocabulary. Therefore, word embedding solves the weaknesses of (BOW) model (Al-Saqqa & Awajan 2019a). The texts that are parallel are represented via similar features. Word embedding is used for multiple NLP applications including sentiment analysis and text classification. In many studies the use of word embedding improved the performance and accuracy of the sentiment analysis. (Al-Saqqa & Awajan 2019a). Although BOW is used widely to analyze sentiment, it still has limitations. For example, the unreliable word order which result in having several documents with same representation since they are using the same words. also BOW overlooks the semantics of words.

Djaballah, Boukhalfa & Boussaid (2019) collected Arabic tweets associated with violence events and labelled them in two labels of sentiments. Their strategy is based on Google's Word2vec scheme, which emphasizes the meaning of words using a deep learning-inspired method. To show tweets, Word2vec and weighted average were utilized. In addition, SVM and Random Forest machine learning techniques were used to measure sentiment. Utilizing the cross-validation technique, we have conducted some studies to verify our methodologies.

Continuous Bag-of-Words Model (CBOW)

In the Continuous Bag-of-Words Model (CBOW) the centre word is predicted based on the nearby text. Then the context is represented by multiple words. by giving s sequence of context words, it will be able to predict the unknown target word according to the window size (Al-Saqqa & Awajan 2019). Djaballah et al., (2019) compared Word2vec with Bag-Of-Words (BOW) approach. They begin by selecting the validation measures to gain more accurate results for the estimation. Next, they evaluate these results by applying the commonly used measures. The researchers tested several state-of-the-art methods that apportioned with the detection and analysis of terrorism-related events. Using sentiment analysis, they classified tweets into two labels and used two approaches Word2vec and Word2vec by weighted average to represent text as vectors. Then, they implemented two main classifiers to predict the outcome. The result of the experiment shows that Word2vec by weighted average have reached better prediction accuracy than using inly Word2vec approach.

Skip-Gram Model (SG)

Skip-Gram Model (SG), predict words in text out to some window size, where each prediction phase uses one word as a focus word. Using this focus word as a starting point, the SG model will provide a probability distribution that represents the likelihood that a word will emerge in the context. The model aims to maximize that possibility distribution for selected vector. CBOW operates a bit quicker than SG to train the model and somewhat higher accuracy for frequent words, on the other hand, SG works better when using smaller amount of training data and represents better on rare word or phrases (Al-Saqqa & Awajan 2019b).

9- Feature extraction

Feature extraction or feature selection comes after the pre-processing stage. Choosing the right features controls the general performance of sentiment analysis. Feature extraction is a technique that allows to eliminate irrelevant, duplicated and noisy data to predict the most related features for the SA approch. Despite that, feature extraction reduces the dimensionality of the feature space and improves the processing time. This results in improving accuracy and efficiency of the model. The most popular Sentiment analysis text feature extraction are n-gramme models and POS. In addition, Tags with stylistic, syntactic, semantic, lexical, and word vector features as well as others can be employed. A contiguous group of n terms from a particular text sequence makes up an n-gramme. The n-gram is referred to depending on the value of the n. Size 1 is designated as a uni-gramme, size 2 as a bi-gramme, and size 3 as a tri-gramme, for instance. The text in which the word (feature) appears determines how much weight it has. There are various types of weighing schemes, including inverse document frequency (IDF) weighting, TF weighting, Boolean weighting, and TF-IDF. SA is considered a challenging task in NLP. Most Arabic Techniques for SA still use expensive hand-made features, whose representation needs manual pre-processing to achieve the desired accuracy. In order to improve their sentiment classification results, researchers found that phrase stemming and POS tagging were quite helpful.

Aydln & Güngör (2021) used two feature extraction methods which are Term frequency and POS tag. Term frequency and specifically tf-idf metric is the frequent occurrence of a word which represents the importance of a word in a piece of text. Additionally, this feature extraction method is broadly implemented in the SA calcification tasks. When the word is discovered frequently it indicates that this word is important and adds meaning to the sentiment. The next feature extraction method is POS tags which has more outstanding statistics compared to the previous tags in SA. Adjectives, for instance, frequently reflect the primary words that convey sentiment. But each word in a POS tag can also be helpful. The POS tags of the words may also be extracted using a variety of morphological assessments and disambiguation methods.

The TF-IDF weight is the creation of dual weights: TF and IDF. the weight of every word in a piece of text are used to compute the weighted average of a feature vector of Tweet a alternative to computing a simple average (Djaballah et al., 2019).

Brahimi et al., (2021). presented a model for extracting sentiments from movie reviews and improving Arabic sentiment analysis. For the task opinion classification, they used role of n-gram and skip-n-gram models. In addition, they applied Part-Of Speech tagging to utilize subjective expressions such as adjectives and nouns. They also added a method to extract relevant sentiments in particular, review that concludes people's opinion. Also phrasing value customer reviews and finding factors that impact their attitudes. The researchers claim that the proposed methods resulted in reaching a 96%acuracy using F-Measure.

10- **Annotation**

Arabic corpus annotation for sentiment analysis includes classifying labels with suitable classes for training machine learning classifiers. Baly et al., (2017) applied Arabic Language Sentiment Analysis (ALSA). It helps in comprehending Arabic sentences on a variety of levels, including phonetically, rhetorically, and metaphorically. The process of labeling acquired data from a corpus collection for ML application is called annotation. Different kinds of corpus annotation exist. For instance, there is the automatic strategy, which makes use of an annotation tool, and the manual approach, which relies on human labor. The categories utilized for the annotation determine how the sentiment analysis will turn out. The Arabic annotation concerns to two levels, sentence level and word level. The annotation can be accomplished manually by respective of native speakers, experts or can be done automatically. According to Baly et al., (2017) Building a high-quality annotated Arabic corpus for sentiment analysis is clearly needed in order to improve classifiers and address research problems while taking into account the six levels of Arabic sentiment analysis.

In a work done by Al-Twairesh et al. (2017) they describe the procedures followed in order to gather and create a substantial collection of Arabic tweets. It is described how to pre-process and clean the collected dataset. A corpus of Arabic tweets with sentiment annotations was taken from this dataset. The vast majority of the corpus is made up of tweets in MSA and the Saudi dialect. The corpus was manually sentiment annotated. An hour-long training session was given to the annotators, who also got annotation guidelines.

They provided a list of rules and explained why each one is important to the annotation process.

(1) News: News is just not recognized as subjective. So, regardless of whether they are delivering good or negative news, they should be marked as neutral.

(2) Perspective: The author's perspective should be used to evaluate the sentiment, and not the annotators.

(3) Context: The label selection should be made considering the text's overall context.

(4) Ambiguity: If the opinion is unclear, they should not make any educated guess; just select the decision (indeterminate).

(5) Mixed: The mixed label must be selected with careful understanding. For example, if a happy emoji appears in a tweet but the sentiment is negative or the opposite, you should select mixed.

CHAPTER 4 DATA COLLECTION

In recent years, there has been a massive growth of data via the usage of mobile apps, internet and social media portals. Due to the increasing development of technologies and social media platforms, people can express their opinion about certain aspects. Many of these social media platforms are heavily used by people around the world to give reviews and opinion on movies. By using technologies such as machine learning (ML) deep learning (DL), it made it easier for people to find movies that meets their expectations by analyzing previous reviews. Moreover, movie reviews have become a vital factor affecting revenues and rating of a movie. Also, movie creators can benefit from these reviews to improve their show based on reviewers' evaluation. However, due to the massive number of reviews and information collected, humans cannot process all data to decide whether a people's opinion is positive or negative. In this case we need to relay on the new ML technologies to have better perdition and analysis for movie reviews.

In this section, we describe in detail for the proposed methodology of data collection. Arabic social media platforms contain of a combination all the three types of Arabic language which CA, MSA, and DA. Although there are similarities between the three types, there are some major differences between the three varieties which can result in poor performance when applying SA. One of the most adaptable and well-liked methods used in data analytics, particularly for machine learning, is the Python language and its libraries. There are many ways to annotate corpora, such as the manual method, which depends on human labor, or the automatic method, which makes use of an annotation tool.

4.1 Training dataset

In this paper we experiment with two types of datasets. Before applying sentiment analysis, we required to train the model and generate a readable version for the machine via dataset annotation (Almuqren & Cristea 2021). Annotation is the task of assigning explanatory

information to a corpus collection for NLP use. Depending on the goal of the annotators, they generally use different annotation schemes. For example, it can have two calcification labels (positive or negative) annotation, or it can have three classification labels (positive, negative, and neural) (Rekik et al. 2018). On the other hand, other annotators can use a number range of numbers, for example 1 to 10 where 1 is strongly positive and 10 is strongly negative.

For the training dataset we used two publicly available data. The first one was retrieved from Kaggle which was collected from twitter using API on Mar 17, 2019. The data set consists of 154415 tweets annotated positive and negative. The second dataset is also available publicly from GitHub and which was collected from twitter API on Mar 17, 2018. This dataset consists of 16,702 tweets and is annotated positive negative and neural. Figure 1 illustrates statistics of the training datasets

Alharbi et al. (2020) presented a novel large Twitter-based benchmark Arabic Sentiment Analysis Dataset (ASAD). This Arabic corpus can be used for general sentiment analysis. The tweets were randomly collected from numerous tweets that were collected by using the Twitter public streaming API6 in the time stamp of May 2012 and April 2020. The dataset has been pre-processed which includes text cleaning such as removing URLs, images, or videos. Therefore, a list of predefined inappropriate Arabic keywords was used for the cleaning task. They used this list to remove spams or any comment that has bad Arabic language. Spam tweets can still be found in the dataset and have an impact on how well a sentiment classifier predicts the future. The finalized corpus includes a total of 100k tweets. By analyzing the content of ASAD, we can summarize that about 69% of the tweets were chosen from the year 2020, 30% were selected from the year 2019 and the remaining 1% were selected from 2012 till 2018. They implemented a three-way classification sentiment. Therefore, a tweet can either be positive, negative or neutral. ASAD has a total of 15,282 positive tweets, 15,349 negative tweets, and 69,369 neutral tweets. Training dataset information are presented in table1.

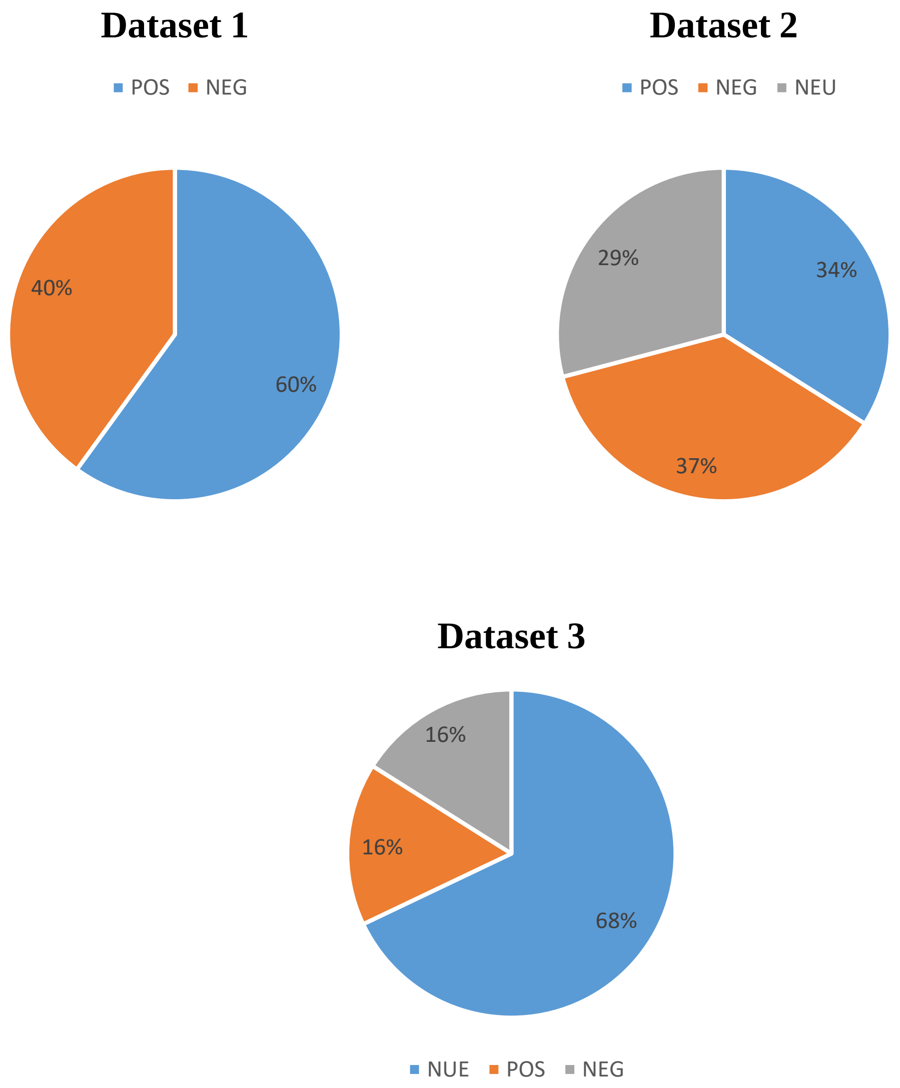


Figure 1: Statistics of the Collected Reviews

Dataset	labels	Date collected	Number of tweets
Dataset 1	Pos, neg	Mar 17, 2019	154,415
Dataset 2	Pos, neg and neu	Mar 17, 2018	16,702
Dataset 3 (ASAD)	Pos, neg and neu	May 2012 and April 2020	100,000

Table 1: Training dataset information

For our experiment We applied deep learning models which are CNN, LSTM, CNN-LSTM, and BIRT model. The three chosen datasets were used as the input for all of these models. It utilizes word embeddings generated from a pretrained word2vec CBOW model. We applied BERT model to Dataset 2 and Dataset 3 because it has three classification labels. However, due to some limitations we did not use the full ASAD dataset, we optimized 55K tweets for the experiment.

4.2 Test data

In this section we will describe the dataset collection methodology. The testing datasets were collected from mainly from Instagram for two Arabic movies reviews. The two movies are:

- **واحد تاني** Wahed Tani which translates to (someone else)
- **عمهم** Amahom which translate to (their uncle)

In general, the methodology involves of three primary stages: collecting data from Instagram, manual annotation and data cleaning. Each stage is further explained in the sections below. Figure3 bellow illustrates the test dataset Collection Methodology.

For the test dataset we harvested two datasets related to the two movies selected. Table 2 includes test dataset information. Movie reviews were collected at two different time steps as follows:

	Movie Name	English translation	Review Date	Number of reviews
movie 1	واحد تاني	Wahed Tani (someone else)	From 25/4/2022 to 11/5/2022	4015
movie 2	عمهم	Amahom (their uncle)	From 23/6/2022 to 24/7/2022	1578

Table 2:Test dataset information

The main lead for the first movie is Amed Helme, therefore, the reviews were collected from his social media sites and mainly from Instagram. The actor has so many posts advertising the movie where people start to comment and give their opinion about the movie. These comments were collected using the tool Vicinitas for Twitter and IGCommentExport for Instagram. Similar steps were done for the second movie were the lead for the is Mohammed imam therefore the comments were collected from his social mead.

Below in figure 2 is the word cloud for the two collected datasets:

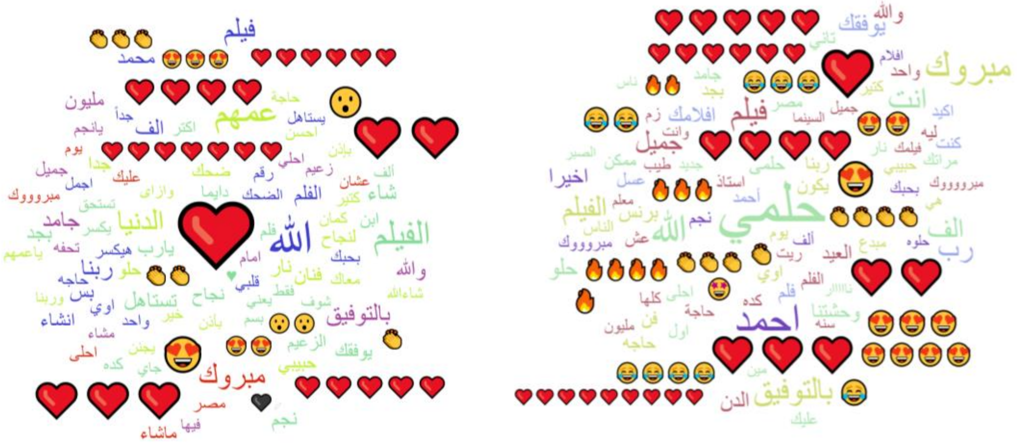


Figure 2:Word cloud presentation of the text found in the reviews of the selected movies

CHAPTER 5 METHODOLOGY

Over the last years, social media sites have gained more popularity and divers. It contains a variety of data components that are in structured different formats such as text, videos, and audios. People are using social media to express their feelings and ideas about most aspects of life. Social media have entered our lifestyle delivering easy to use platform people to express their own thoughts and exchange information worldwide. For example, people comment on social media to typically express their opinion about a movie they saw, or restaurant they visited or hotel. Massive data can be collected every day from social media networks such as YouTube, twitter, Instagram, and many other platforms. There are different ways which these data can be collected depending on the nature of the platform. For social media most studies are done in English and less done on Arabic. In this paper we emphasize collecting only Arabic sentiments while taking into account the characteristics of social media platforms and other factors. Nevertheless, Arabic data is collected from different social media sites using two different tools and stored on a local database folder. Our approach focuses on studying Arabic text with emojis which can add value to the meaning and emotions of the user.

SA is a popular and difficult NLP process that makes it easier to gather public opinion on a subject, good or service. As a result, there are numerous benchmark datasets and algorithms available to address this difficulty in this heavily researched study area.

In tis paper we will be applying deep learning SA on Dialect Arabic language. We are also using an Arabic version of BERT which was also trained on DA. The use of multiple dialects in the dataset makes a challenge to the SA model.

Since these terms vary from dialect to dialect, the range of keywords linked with each sentiment expands when there are numerous dialects. This adds a new difficulty that the corpus of a single dialect does not have. The multi-dialect dataset may not respond well to a pre-trained word embedding model with a one-dialect corpus.

BERT is one of the recent language representation algorithms created by Google (Devlin et al. 2019). BERT can remain adjusted by one further output layer to produce a new model for a certain NLP task. Many bilingual BERT models are now published and can support various languages. AraBERT is one of the Arabic language models which are trained on MSA. Using a specified Arabic corpus trained BERT model outperforms the performance of a multilingual version of BERT. The model is trained on a huge-scale Arabic dataset and can be used with many NLP tasks including SA. Abdul-Mageed et al., (2020) introduced two powerful Arabic deep bidirectional transformer-based models ARBERT and MARBERT. ARBERT was pre-trained on 61GB of MSA and MARBERT is centered on both Dialectal Arabic (DA) and MSA. Figure3 illustrates the test dataset Collection Methodology.

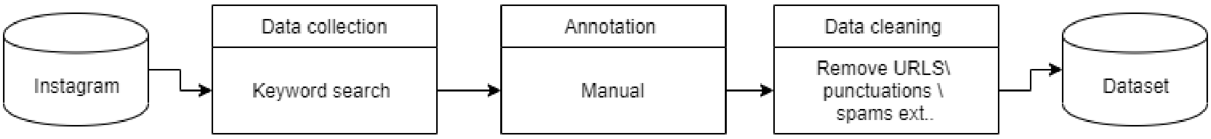


Figure 3: Dataset Collection Methodology

5.1 Tools used for dataset collection

The tool used for twitter corpus collection is a web-based tool called Vicinitas. Which allows to collect a stream of a historical tweets and real-time tweets containing selected hashtags, keywords, and user mentions. This tool is ideal for analyzing old tweets related to a hashtag, keyword or account since 2006 until today. The free version allows to collect up to 2000 tweets in the last 7 days. The upgraded version of historical tweets allows us to collect a large amount of data for the selected period of time. Real time tweets are collected from the present moment to the upcoming/future tweets not as the historical tweets which are prior to starting the tracker. Vicinitas begins collecting tweets related to your tracker about 10 minutes after you initialized. The data then can be exported to an excel file which allows for data to be cleaned and applying pre-processing techniques.

Another tool used for collecting Instagram comments is called IGCommentExport. It can be downloaded as an extension for google chrome and allows to download up to 10000 comments at a time and export it as a csv format.

There can be a mismatch between the number of comments on the post and the number of comments displayed in your file when you export comments from an Instagram post. These reasons are the most likely ones. First according to the tool’s official website, a private user's comments cannot be exported. Due to their privacy settings,

You may be able to see their comment on the post but, you will not be able to export it. The second reason is that the post is labels as was advertisement or promotion. Also, a user's comment won't be pulled in if it was made on a post that was displayed to them as an advertisement. Fields that provide aggregated values will not include ads-driven data, according to Facebook/IG. For instance, comments count counts comments on photos but not on ads that feature those photos.

5.2 Exporting data

We can export the specifics of each of the most recent up to 2000 tweets associated with a tracker into Excel for further analysis and use. The Excel file would include fine-grained details about each tweet, including its text, its author, how many likes and retweets it has received, its category, whether it includes rich media, when it was posted, and more. For Instagram the csv file was also converted to an excel file which contains information as username, time posted and comments.

5.3 Data Collection challenges

Collecting data from twitter contained a lot of irrelevant information such as advertisements, malicious profile, and cyber bulling. To avoid getting these unnecessary data We adopt a keyword-based approach which focuses on retrieving data related to the predefined keywords. We develop a set of search phrases to improve the likelihood of finding tweets that express thoughts, attitudes, or feelings about the given entities in order to get tweets that are pertinent for the chosen movie. The keywords used are

(Ahmad_helmy, Wahed_tani, Amahom_movie, Mohamed_Imam_Amahom, Amahom)

(احمد_حلمي, واحد_تاني, فيلم_واحد_تاني, عمهم_فيلم, محمد_امام_عمهم, عمهم)

Using filters created on the online content is another option. Twitter commonly uses media content, such as photographs, videos, and GIFs, which typically receive a lot of engagement in the form of likes and retweets. Therefore, we can use filters to parse accurate content by benefiting from photos, videos and retweets. In addition, we may group every recovered tweet relying on when it was posted or how many likes and retweets it earned.

For Instagram tool, there is a difference between the number of comments shown on the post compared to the number of comments shown the dashboard. This is due to the fact the tool (IGCommentExport) will not include ads-driven data from posted comments.

5.4 pre-processing:

Preprocessing is an essential data preparation phase for sentiment analysis. Data preprocessing is an important phase in NLP tasks to improve the performance of the dataset and ensure the reliability of the sentimental analysis (Alkhair et al. 2019).To perform the preprocessing first the data was cleaned manually and collected in a single excel file saved locale on the computer. By using excel we removed the unnecessary data to make them ready when required to be used for sentiment analysis. In the cleaning process unwanted data such as, URLs, punctuations, usernames (mentions), and numbers Alayba et al., (2018). By analyzing the content of comments, we can find some spam which were then removed manually. A lot of comments were using emojis to express the reviewer’s opinion. Some comments also can include only emojis. Therefor the emojis were translated to Arabic words related the meaning of each emoji. We also call this process Emoticons which then will replace emojis with universal happy, sad tokens

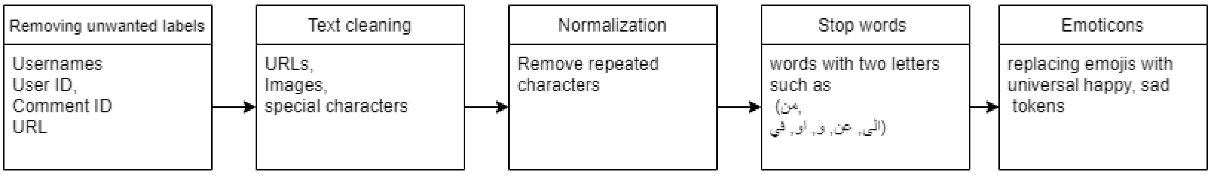


Figure 4: Preprocessing Methodology

In this paper the dataset preprocessing step was as follows, we applied normalization and data cleaning to dataset as illustrated in the figure 4:

- Removing unwanted labels such as, usernames, user ID, comment ID and URL
- Eliminating all English or none Arabic words.
- Removing all digit numbers
- Removing special characters while keeping the characters related to emojis since its describes and provides value the writers emotions. The characters removed were as (!@#$%^&) since it does not add value to the sentiment.
- normalization for example the letters (وٌ ؤ وّ) is converted to the common letter و.
- Normalizing repeated characters which is also known as word elongation
- Emoticons by replacing emojis with universal happy, sad tokens
- Stemming: As an example, the word wait, waits, waiting and waited all become wait when stemmed. Or in Arabic, the root of the word teachers (**معلمات**) is (**علم**) and the stem is (**معلم**)**.**
- Stop words removal: is by removing words that do not contribute to the word meaning or add any value to the context. Mostly these words are used for language structuring. As an example, are a, are, the, was, ext. In Arabic we remove word that has only two letters such as (**من, الى, عن, و, او, في**)
- vectorization: The reviews written in social media, or any other sources are written in text format which humans can easily understand. However, neural networks currently are unable to understand raw text format and cannot be used directly with the given text data. Pre-processing needs to be applied to convert the data to a format that neural networks can understand which is vector format. The text needs to be converted to a numerical format called vector. The text can be utilized as an input for the rebuilt model once it has been converted to vector form. Vectorization of the written material is the process of transforming text into a vector. (Bodapati, Veeranjaneyulu & Shaik 2019c)

5.4.1 Emojis Translation

We translated some of the most common emojis as per their meaning in Arabic. according to Tomihira et al., (2020) over the last few years Emojis have been standardized and are now commonly across social media platforms. It can be sometime that the expression of a sentence can have a neutral expressive state. So then, adding an emoji can easily emphasize the emotion and a sentiment to this sentence.

Emojis could be utilized in this scenario to heighten the writer's feelings. In other words, it enables the author to clarify an emotional passage by emphasizing one particular feeling while a range of emotions may be present. The inclusion of an emoji gives one of the two emotions happiness or sadness priority. Below in table 3 is the emoji translation table.

Emoji	Translation Arabic	Translation English	Label
♥ 💕 💝	قلب	heart	pos
😂 🤣	يضحك	laughter	pos
🥰 ☺ 🙂	مبتسم	smile	pos
😍 😁	سعيد	happy	pos
😘 💋	قبلة	kiss	pos
🥺 😔 ☹	حزين	sad	neg
😭 😢	يبكي	cry	neg

🤦‍♀️ 😅 🙃 😐	متعجب	wonder	nue
🎉 🥳	حفلة	party	pos
😱	يصرخ	Scream	neg
🙏	شكرا	thank	nue
👍 👌	موفق	ok	pos
🙈	محرج	Embarrassed	neg
🔥	نار	fire	pos

Table 3: Emoji translation

5.4.2 Annotation

Before applying sentiment analysis, we have to create a readable version of the data for the machine in order to train the ML algorithms using annotation. Annotation is the procedure of labelling collected data of a corpus collection for ML use. There are several types of corpus annotation. For instance, there is the automatic strategy, which makes use of an annotation program, and the manual method, which relies on human labor. The results of the sentiment analysis are relying on the categories used for the annotation

5.4.3 Google News Word2Vec

The Arabic word representation (Embeddings) is available free for research use. It uses Mikolov's Skip-gram and CBOW models. It is an extension to Mikolov's toolkit. Word Embedding is a mathematical model representing a word in space via a vector. Every dimension is a feature to the represented word. The feature can have semantic or syntactic definition. the connection among words will be measured by means of a similarity function. The word is then mapped to a numerical representation. Google News Word2Vec is an effective method for generating word vector representations using the continuous bag-of-words and skip-gram architecture. Word vectors are produced by the word2vec tool using a text dataset as its input. From the training textual data, a vocabulary is first created, and then words are represented as vectors. Numerous NLP and ML applications can infer features from the produced word vector. Finding the closest words to a user-specified term is a simple way to test the representations that have been learned. In this instance, the distance tool is applied to that end.

5.5 Challenges in processing Arabic text

The characteristics and complications of the Arabic language have made processing Arabic sentiments, a main challenge (Al-Twairesh et al. 2017). These characteristics need to be considered when applying any sentiment analysis to Arabic text. Researchers dealt with major difficulties such as ambiguity, diglossia, morphology, and general understanding of the Arabic text. For example, it has complex rules and grammar where the meaning of a word is dependent on its root and overall, the full syntax. According to Rieser & Refaee (2014) The intricate nature of Arabic's morphology, structure, and grammar is one potential explanation for the language's complexity. The nature of the language allows it to have rich dialects and vast numbers of synonyms. Also, words in Arabic have no capitalization or dedicated letter. normalization is inconsistent when used with certain words, letters or diacritical marks. Therefore, a lot of pre-processing steps are needed when dealing with the Arabic context.

CHAPTER 6 EVALUATION AND RESULT

The three datasets were applied to the proposed CNN, RNN, CNN-RNN, and BERT models. Two of these datasets were used with Bert model and compared with the other tested models. We first tested the CNN model then, the LSTM than, combination of CNN-LSTM. These three modes were compared then the best mode was selected to further compare it with the BERT model. The datasets are generally pre-processed to improve the quality of the training dataset. We applied tokenization to distinct numbers and punctuation from words. Emoticons were applied to replace emojis with words such as in the table 3. Each dataset is split into 80% training and 20% validation. Then the models were used on the testing dataset of selected two movies.

Recent studies have proven that twitter is one of the major sites to obtaining information about perceptions and views on an event or product (Hasan et al., 2018). Many researchers proposed different algorithm for manipulating the data and extracting emotions from tweets. SA can achieve great results when used with a large scale of data. Therefore, many SA researchers built their corpus using twitter data. In this paper we used data written on both twitter and Instagram regarding a reviewed Arabic film. The two films are, واحد تاني (someone else) and عمهم (their uncle).

The experiment was done using python. We start the experiment by downloading the required libraries. Which mainly are numpy, pandas, os, collections, string, and nltk. Then we import the selected training dataset for the model to read it. A set of pre-processing were applied to the data such as text cleaning, removing stop-word and tokenization. The training data is split to training and evaluation data. Each selected dataset is split into training (80%) and testing (20%) while maintaining similar sentiment distributions giving subsets. Google News Word2Vec is used and where the data have been trained first to get the vectors. Next is to define the CNN model as presented. This mode uses a number of five conventional layers. After training the mode we evaluate the results to make sure it will predict the movies correctly. In this sentiment analysis the data is labelled into two classes which are positive and negative. Finally provide the model with the movie dataset for the prediction process. The second experiment is applying SA using CNN and using Google News Word2Vec.

The CNN framework used is a type of deep learning neural network. Convolutional neural networks are a mathematical function that stimulates human neural networks. The presented CNN consist of five Convolutional layers in addition to the input and output layers. Figure 5 shows a graphical representation of the proposed CNN network. It represents the model framework for sentiment analysis. The network consists of an embedding layer, 5 convolutional layers, a pooling layer, and a fully connected layer.

6.1 Results of CNN model

First the CNN model was tested on a dataset that has two classification labels which are positive and negative. The accuracy reached 91% which is a greater result. however, when using the same parameter with a different dataset that has three classification labels positive, negative, and neutral, the model achieved lower accuracy of 46%. Therefore, to increase the accuracy we increased the number of epochs from 3 to 5 which then achieved the accuracy of 65%. The results show that CNN model was able to predict a number of 598 positive, 591 negative and 485 neutral using dataset 2. More accuracy can be reached on adding more epoch

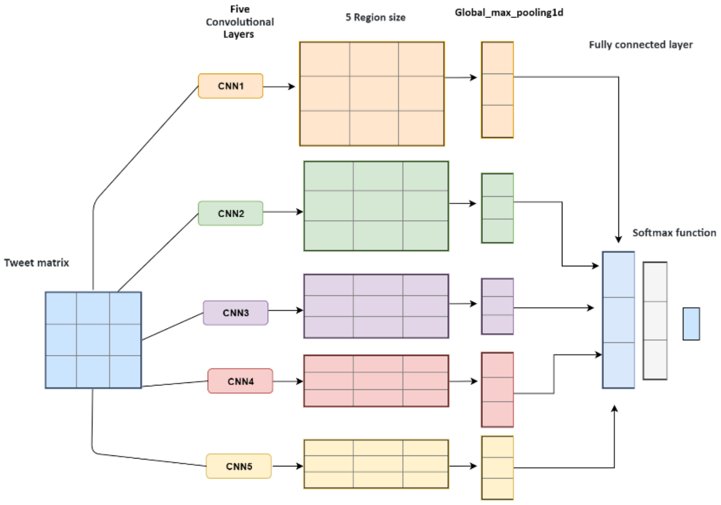


Figure 5: CNN model Architecture

6.2 Result of LSTM model

RNN -LSTM model was used with the dataset that has two labels. The results show that it achieved a similar result for accuracy 91% in epoch 3 however there is a slight increase in accuracy which reached 92% on epoch 5. Results direct a slight advantage of RNN-LSTM over CNN by looking at the LSTM outcomes. Figure 6 shows a graphical representation of the proposed LSTM network

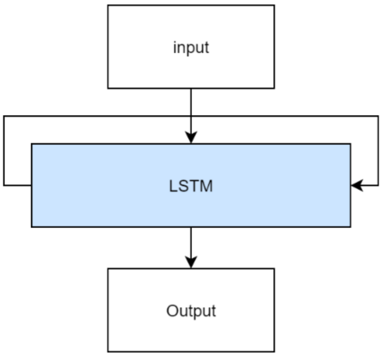


Figure 6: LSTM Model Architecture

6.3 Result of CNN-LSTM model

Next, we applied a combined CNN-LSTM model using the same 3 class label dataset. This model used four layers of CNN in addition to LSTM layer. The results showed an accuracy of 90%. Compared to the previous CNN and RNN model it achieved better results with the same number of epochs 5. Finally, we perceive that both CNN and LSTM models performed knowingly better on the two-class dataset while it requires more learning when applied on three class datasets.

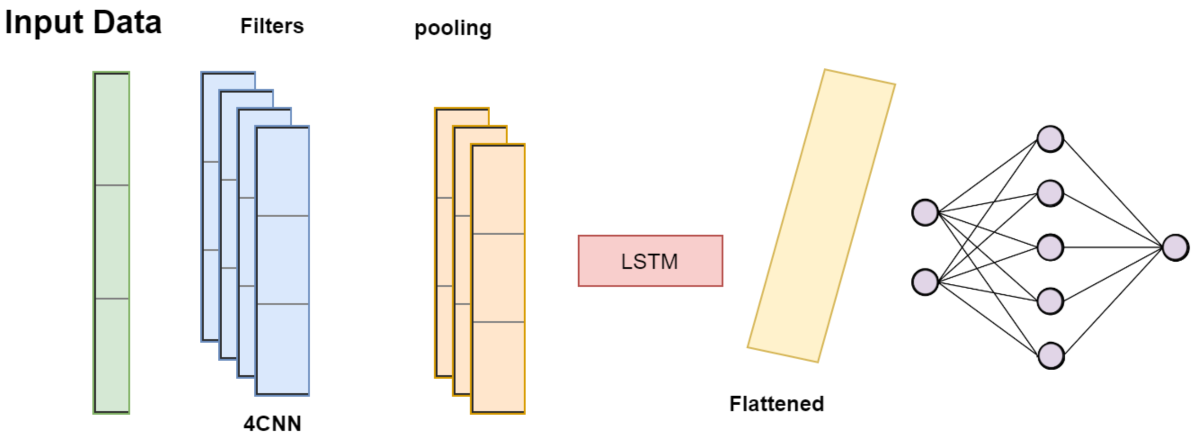


Figure 7: CNN- LSTM Model Architecture

6.4 Result of BERT model

As a result, we applied the pre-training and fine-tuning stages as stated below, adhering to the previously reported training process. This experiment is conducted by implementing BERT on Arabic Sentiment Analysis (ASA) using Python. The collected datasets as mentioned before will be used in this experiment. Dataset 3 is the ones used to train the model. The results are compared with the CNN-LSTM since it is the best performing model among the previous models. Therefore, we will be testing two of the latest state of art models. The first experiment was processed using CNN-LSTM model while the second was processed using BERT model that was specifically built for Arabic called MARBERT. MARBERT is a large-scale pre-trained masked language model focused on both Dialectal Arabic (DA) and MSA (Abdul-Mageed et al., 2020). The dataset 3 was split into three subsets: 80% for training, 10% validation, and 10% for testing. Figure 8 shows a graphical representation of the proposed BERT network where it represents the model framework for sentiment analysis.

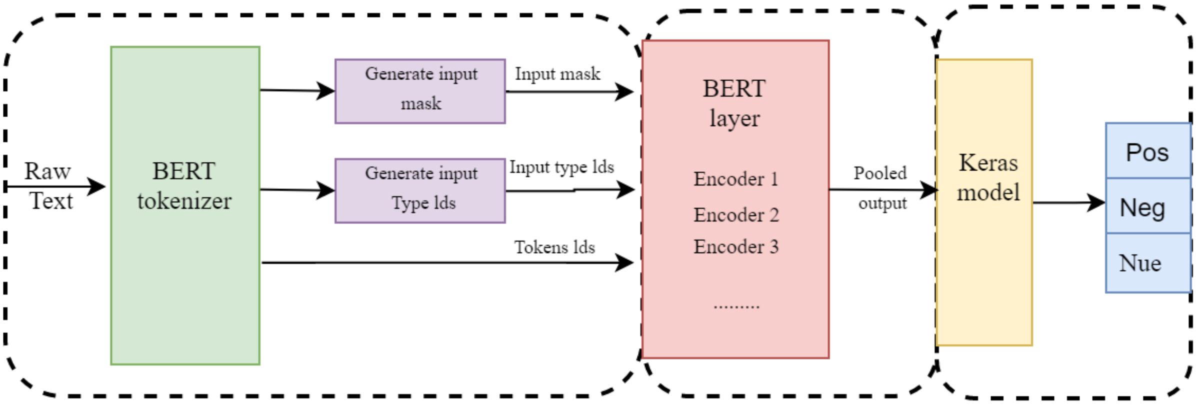


Figure 8: BERT Model Architecture

We are using python language in the implementations and Jupyter Notebook that support the ML and data science projects. After training on the BERT model, we will evaluate the model to check the performance of trained model and will select the final model for classification of Arabic tweet sentiments.

In this paper we used 55K annotated tweets from dataset 3 (ASAD) with three-class sentiments. The dataset originally includes 100K annotated tweets for Arabic sentiment analysis. A total of 69,369 neutral tweets, 15,282 neutral tweets, and 15,349 neutral tweets have been posted by ASAD. Between May 2012 and April 2020, the tweets were gathered. They are written in a variety of Arabic dialects, namely DA, MSA, Egyptian, Khaleeji, and Hijazi.

In this paper, we used an Arabic version of the BERT model: MARBERT that uses both Dialectal Arabic (DA) and MSA, it is pre-trained from scratch and made publicly available for use. The main references for this BERT model are (Abdul-Mageed et al., 2020; Antoun et al., 2020). This model is extensively considered as the base for most state-of-the-art results in different NLP tasks (Antoun et al., 2020). Our method is similar to that of Abdul-Mageed et al. (2020), who used a BERT-base design with 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and 110M parameters overall. In this experiment, the original application of pretraining BERT on transformer model was used. We start by Installing the needed packages and importing them to the model. The main were pyarabic, emoji, pystemmer, optuna==2.3.0 and transformers==4.2.1. PyArabic is A designated Arabic language library for Python, which delivers fundamental functions to handle Arabic letters and text. For example, it detects Arabic letters, Arabic letters sets and characteristics, removes diacritics etc. To prepare the raw training data for the model, we applied light pre-processing. In order to better fit the Arabic language, we perform additional pre-processing before the BERT model's pre-training. After the pre-processing we split the training Dataset to train and evaluation as seen in the table 4 below.

Split	Neutral	Positive	Negative
Training Set	50481	11623	11436
Evaluation Set	46	14	14

Table 4: Training and evaluation set

Then we start preparing the model by setting the max_len. This step will be beneficial for the BERT Model. We can use extra padding in length which is initiated to be beneficial for increasing F-score. After that we start preparing BERT Model Classes. Then defining Needed Methods for training and evaluation. Then Define Training Arguments and build the trainer. Next step is to apply the trainer to train the model. After training the model an evaluation is performed to test the reliability of the model. Finally, we apply the model to the test dataset to predict the labels for the sentiments.

Pre-processing for BERT:

The preprocessing was performed for the selected dataset 3 before applying the model.

The additional preprocessing step which essentially involve eliminating punctuation, Arabic diacritics (short vowels and tashkeel), elongation, and stop-words. Some of which are available in NLTK library. This helps in maintaining an accurate representation of the original written text. The pre-processing steps were similarly applied on both training and test set. For text pre-processing pyarabic and pystemmer were used. Multiple techniques for calculating a word's "stemmed" form are accessible through PyStemmer. The majority of the typical morphological ends have been eliminated from this form. In Arabic it provides the root if the word; expectantly representing a common linguistic base form. Pyarabic is a library that includes features such as Arabic letters classification, Harakat Strip, Reduce tashkeel and Letters normalization. Also the normalization stage helps in replacing letter such as (آ,إ,أ) with more simple letter (ا). The preprocessing results in reducing data size and enhancing prediction efficiency. A lest of stop word was retrieved and used to remove Arabic stop words.

Result:

The results in Table 7 show that BERT model performed well on the Arabic language on most tested datasets. Even though BERT was trained on ASAD dataset, the model was able to predict sentiment on texts that were never seen before. The experiment began by building a baseline model using a deep BERT model approach for sentiment analysis. In addition, the selected BERT model was compared with further recent CNN_LSTM approaches. The effectiveness of the tested model can be assessed using a variety of techniques. One of the popular approaches for evaluation is accuracy, which we utilized to assess the effectiveness of our experiment. Using only one epoch, the BERT model scored 91% accurate. The measures used are precision, recall, f1-score and support which are shown in the table 8 below.

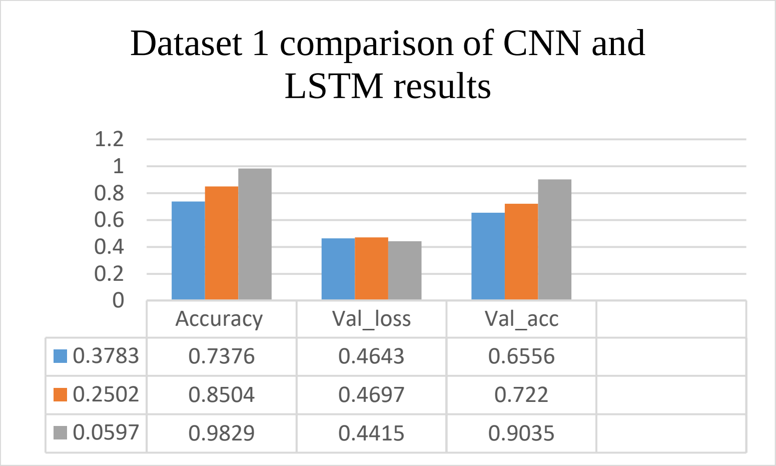
Finally the model will be used to predict the test datasets by using Python predict() function. This function allows us to predict the labels of the input data by maintaining the trained model. The predict() function receives only a single argument which is generally the data to be predicted. It analyses the labels of the data based on the trained data gained from the model. Thus, it works within top of the learned model and benefits from the learned label to map and predict the classes of labels for the data to be tested. In our experiment the classes are neutral, positive and negative

6.5 Comparison of results:

Knowing the model accuracy is significant since it allows us to evaluate how accurately a given model will predict test data. After conducting several experiments using CNN parameters on three datasets, the results are shown below. Results indicate a clear advantage of RNN-LSTM over CNN. The LSTM results show that abstracting away from words improves performance, particularly when lemmas are included, which is consistent with the findings in (Antoun et al., 2020). Finally, we can observe that CNN models performed significantly better on the two class (Dataset 1). CNN uses a similar structure as the previous model, but without pre-trained BERT as an embedder.

The CNN and LSTM algorithms performed very well on the first dataset that has two class labels. With a slight advantage for the LSTM reaching 0.907 accuracy and a little more improvement when adding 5 epochs. Below in figure 9 is the comparison of the two models CNN and LSTM tested.

Dataset 1

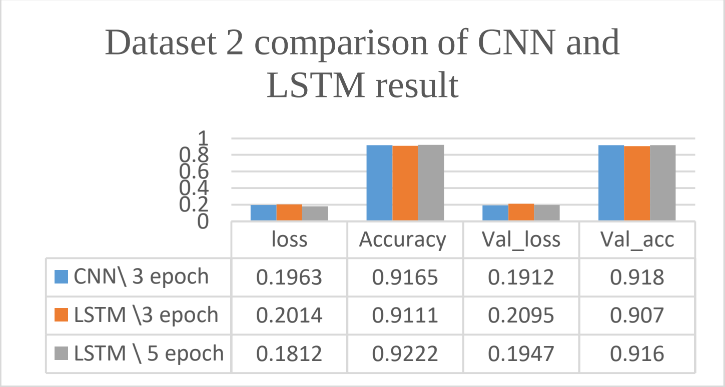


	Accuracy	Val_loss	Val_acc	
🟦 0.3783	0.7376	0.4643	0.6556	
🟧 0.2502	0.8504	0.4697	0.722	
⬜ 0.0597	0.9829	0.4415	0.9035	

Figure 9: Dataset 1 comparison of CNN and LSTM result

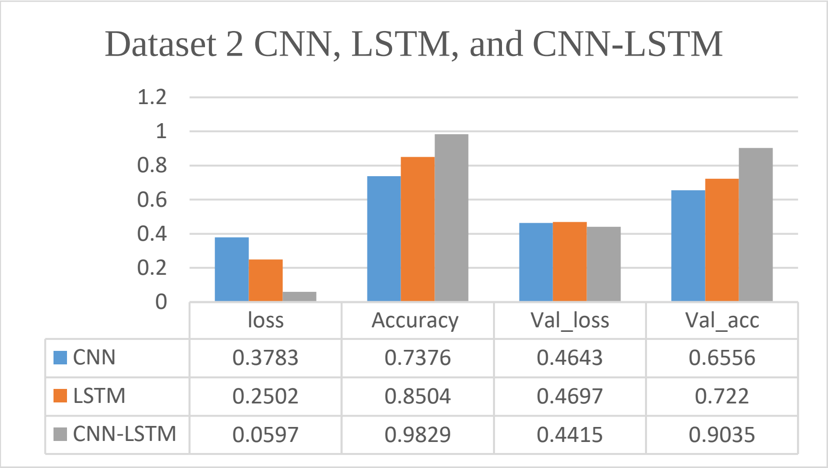
Next the same CNN and LSTM models were tested on the second dataset that has a three-class label. Compared to the previous dataset, the CNN did not perform well on the second dataset which reached the accuracy of 0.4579. To get a better result we increased the epoch to a had slightly more improvement of 0.6656. Next, we used the LSTM model which had a somewhat better result than the CNN which reached 0.7220. Finally, we applied a combination of CNN-LSTM which achieved a greater result reaching 0.907 accuracy. We conclude that combining the two models achieved better results for sentiment analysis. Figure 10 illustrates the results of Dataset 2 with comparison of CNN and LSTM result. Figure 11: illustrates the results Dataset 2 with comparison the three models CNN, LSTM, and CNN-LSTM

Dataset 2



	loss	Accuracy	Val_loss	Val_acc
🟦 CNN\ 3 epoch	0.1963	0.9165	0.1912	0.918
🟧 LSTM \3 epoch	0.2014	0.9111	0.2095	0.907
⬜ LSTM \ 5 epoch	0.1812	0.9222	0.1947	0.916

Figure 10: Dataset 2 comparison of CNN and LSTM result



	loss	Accuracy	Val_loss	Val_acc
🟦 CNN	0.3783	0.7376	0.4643	0.6556
🟧 LSTM	0.2502	0.8504	0.4697	0.722
⬜ CNN-LSTM	0.0597	0.9829	0.4415	0.9035

Figure 11: Dataset 2 result comparison CNN, LSTM, and CNN-LSTM

Since the CNN-LSTM achieved the best result, we compared it with the New Arabic BERT model. Note that, two datasets we tested in this experiment. The dataset 3 achieved superior result with just one epoch reaching 91% accuracy. A similar result was achieved using Dataset 2. According to a similar test done by Safaya et al., (2020) We can visibly see the improvement that was made by BERT compared to CNN by looking at the average outcomes of the BERT model on its own. Additionally, we can see that employing Arabic-specific pre-trained models is advantageous compared to utilizing a multilingual model.

6.6 Evaluation metric

For the evaluation section we adopt accuracy, precision, recall, and F1 score as the evaluation metrics to compare each method's overall classification performance across three classes: neutral, adverse, and positive Performance was evaluated using the following four key metrics:

Accuracy is defined as the number accurate forecasts divided the number predictions. Accuracy displays the proportion of texts that were successfully classified.

Precision is defined as the proportion of accurate positive outcomes to all positive results. The accuracy of the positive class prediction increases with the precision percentage.

Recall is calculated by dividing the total number of expected positive outcomes by the correctly predicted positive results.

F1-score is the average of precision and recall, with 1 being the best and 0 being the worst.

The results for Dataset 2 and dataset 3 using the hybrid CNN-LSTM are shown on the table 5 below. Where we can see that the accuracy of the dataset 2 was significantly higher achieving 90%. However, when used with dataset 3 it scores much lower reaching 73%.

Measure	Precision	Recall	F-Measure	Support	Accuracy
Dataset 2	0.96	0.89	0.93	1239	0.90
Dataset 3	0.44	0.52	0.48	1764	0.73

Table 5: Comparing Dataset 2 and Dataset 3 Evaluation metric using CNN-LSTM

The results for BERT model and CNN-LSTM using Dataset 3 are compared in the table 6 below. We can clearly see the BERT model performed better than the CNN-LSTM reaching an accuracy of 91%

Measure	Precision	Recall	F-Measure	Support	Accuracy
CNN-LSTM	0.44	0.52	0.48	1764	0.73
BERT	0.83	0.71	0.77	14	0.91

Table 6: Dataset 3 CNN-LSTM and BERT Evaluation metric

We compared the outcomes obtained while applying several DL models to the three class datasets after under sampling based on the experimental results presented in the table 6. The results revealed that both BERT and CNN-LSTM achieved grate results reaching 91% for BERT and 90% for CNN-LSTM. However, we should note that the CNN-LSTM used 5 epochs and BERT 2 epochs. Therefore, the BERT model outperformed all other classifiers in terms of accuracy (91%), recall (71%%), precision (83%), and F-measure (77%).

On the other hand, the least performance in terms of accuracy, was achieved using CNN classifier when applied on the 3 classes datasets which are (dataset2 and dataset3). It is worth noting that performance CNN and LSTM separately are good when applied on two class datasets. It is also possible to increase the accuracy performance when adding more epochs.

6.7 Answers to the research questions

RQ1 How dose different deep learning models perform on different dialect Arabic datasets?

We can see that the models performed differently depending on the number of classes. For example, the CNN and the LSTM models performed well on dataset 1, which had two classes. However, when applied on dataset 2, which had 3 classes it required more learning.

RQ2 What is the best performing tested model for SA?

We tested four different models which are CNN, LSTM, a hybrid CNN-LSTM, and BERT.

The best performing models are hybrid CNN-LSTM, and BERT which produced high accuracy. A small advantage runs to the BERT model since it required less learning and generated slightly higher accuracy.

RQ3 What is the effect of preprocessing on the quality of Arabic sentiment analysis?

The preprocessing step improved the quality of the training data. We utilized emojis in the prepressing step a translated it to an equivalent word meaning. This prosses was important since the majority of Instagram comments are based on emojis and it adds value to the meaning.

CHAPTER 7 FUTURE WORK

For future work we will apply a combination of CNN with a pre-trained BERT to achieve a better result. The BERT model achieved great results when used on its own. Therefore, (Safaya et al., 2020) achieved superior results combining BERT with CNN. Moreover, we can obviously perceive the benefit of using Language-specific pre-trained models on NLP tasks.

Higher specs are recommended to be used to get better outcomes for BERT model. Because of the pc's limitations the model did not achieve it full potential. For instance, it was suggested that the test be run using GPU and the BERT model. However, training the model on the current PC was time-consuming, which limited the quantity of the training dataset. In addition, the PC's low power and lengthy processing time prevented us from running for more than two epochs.

In addition to that, emojis need more understanding as they are crucial symbols for representing sentiments. In a newer version of ARABERT The two updated models libraries now includes common terms that weren't initially present as well as new vocabulary and emojis. The pre-training was done with a max sentence length of 64 only for 1 epoch.

In our research we did not include the author’s ID. In future research we can include the identity of the author of the reviews. This leads to identifying spam accounts, and one user can comment many times where, it needs to be combined in one comment. This will also need to be considered to prevent fake reviews, which can manipulate the results.

We will also continue finding better structures for the sentiment analysis by adding more layers and combining different models.

CONCLUSION

In conclusion, Sentiment analysis (SA) is a developing field of study in Natural Language Processing (NLP). Automatically detecting attitude or sentiment in a text is often helpful. Sentiment analysis is the automation of the process of identifying and categorizing the reviewer's expressed feelings in a written text on a particular subject or item. It enables mining of the massive amounts of data shared on social networks. People share their ideas on social media, review websites, and blogs. SA's role is to analyze the provided textual data to identify the emotions in it. movie creators can benefit from these reviews to improve their show based on reviewers' evaluation. By using technologies such as machine learning (ML) deep learning (DL), It became simpler for individuals to find movies that live up to their expectations. Humans, however, are unable to evaluate all of the data to determine if a person's opinion is favorable or negative due to the vast number of reviews and information gathered. In this situation, relying on emerging ML technologies will help us produce and analyze movie reviews more effectively. We discovered that artificial intelligence, along with the development of DL and ML techniques, may be used to comprehend and extract important knowledge from the vast amount of data existing on the internet. We came to the conclusion through experimentation that models utilizing neural network approaches, such as CNN and LSTM, more accurately predicted the outcome than models utilizing conventional machine learning techniques, such as SVM, LSTM, Naive Bayes, and KNN.As a result, it can be concluded that using hybrid models, like CNN-LSTM, achieved better results than using CNN on its own, or LSTM on its own. Combining other machine learning approaches provides a more comprehensive and promising results. Emotional analysis allows for a deeper comprehension of the particular emotions underneath those overarching sensations by categorizing the data into whether it exhibits a positive or negative feeling. emotions that are frequently associated with behavior. Pre-processing the data makes it possible to increase the dataset's quality and ensures that sentiment analysis will work more effectively. The prepressing procedure can be completed in a number of steps, depending on the study purpose and the chosen language. Arabic social media platforms contain of a combination all the three types of Arabic language which CA, MSA, and DA. Processing Arabic sentiments has proven to have major challenges because of the features and complexities of the Arabic language. In our experiment, we used two distinct programs for data collection, Arabic data is gathered from several social networking sites and exported into a local database folder. Our method focuses on analyzing Arabic text that includes emojis, which might enhance the user's understanding and feelings.

REFERENCES:

A. Al Shamsi, A. & Abdallah, S. (2022). Sentiment Analysis of Emirati Dialect. *Big Data and Cognitive Computing*. MDPI AG, vol. 6(2), p. 57.

Abdul-Mageed, M., Elmadany, A., Moatez, E. & Nagoudi, B. (2020). ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *Natural Language Processing Lab*, pp. 7088–7105 [online].Available at: https://github.com/attardi/wikiextractor.

Alayba, A. M., Palade, V., England, M. & Iqbal, R. (2018). Improving Sentiment Analysis in Arabic Using Word Representation. IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR) [online].Available at: https://bitbucket.org/a_alayba/arabic-health-services-ahs-dataset/src].

Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I. I. & Zhang, X. (2020). ASAD: A Twitter-based Benchmark Arabic Sentiment Analysis Dataset [online].Available at: http://arxiv.org/abs/2011.00578.

Alkhair, M., Meftouh, K., Othman, N., Smaïli, K., An, K. S. & Smaili, K. (2019). An Arabic Corpus of Fake News: Collection, Analysis and Classification. CCIS, vol. 1108, pp. 292–302.

AlKhatib, M., el Barachi, M., AleAhmad, A., Oroumchian, F. & Shaalan, K. (2020). A sentiment reporting framework for major city events: Case study on the China-United States trade war. *Journal of Cleaner Production*. Elsevier Ltd, vol. 264.

Alkhatib, M., el Barachi, M. & Shaalan, K. (2019). An Arabic social media based framework for incidents and events monitoring in smart cities. *Journal of Cleaner Production*. Elsevier Ltd, vol. 220, pp. 771–785.

Almuqren, L. & Cristea, A. (2021). AraCust: a Saudi Telecom Tweets corpus for sentiment analysis. *PeerJ Computer Science*. PeerJ Inc., vol. 7, pp. 1–30.

Al-Saqqa, S. & Awajan, A. (2019). The Use of Word2vec Model in Sentiment Analysis: A Survey. *ACM International Conference Proceeding Series*. Association for Computing Machinery, pp. 39–43.

Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A. & Al-Ohali, Y. (2017). AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. *Procedia Computer Science*. Elsevier B.V., pp. 63–72.

Antoun, W., Baly, F. & Hajj, H. (2020). *AraBERT: Transformer-based Model for Arabic Language Understanding*.

Assiri, A., Emam, A. & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. *Journal of Information Science*. SAGE Publications Ltd, vol. 44(2), pp. 184–202.

Aydln, C. R. & Güngör, T. (2021). Sentiment analysis in Turkish: Supervised, semi-supervised, and unsupervised techniques. *Natural Language Engineering*. Cambridge University Press, vol. 27(4), pp. 455–483.

Baly, R., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., Shaban, K. B. & El-Hajj, W. (2017). Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects. *Procedia Computer Science*. Elsevier B.V., pp. 266–273.

Barrón Estrada, M. L., Zatarain Cabada, R., Oramas Bustillos, R. & Graff, M. (2020). Opinion mining and emotion recognition applied to learning environments. *Expert Systems with Applications*. Elsevier Ltd, vol. 150.

Bodapati, J. D., Veeranjaneyulu, N. & Shaik, S. (2019). Sentiment analysis from movie reviews using LSTMs. *Ingenierie des Systemes d'Information*. International Information and Engineering Technology Association, vol. 24(1), pp. 125–129.

Brahimi, B., Touahria, M. & Tari, A. (2021). Improving sentiment analysis in Arabic: A combined approach. *Journal of King Saud University - Computer and Information Sciences*. King Saud bin Abdulaziz University, vol. 33(10), pp. 1242–1250.

Chouikhi, H., Chniter, H. & Jarray, F. (2021). Arabic Sentiment Analysis Using BERT Model. *Communications in Computer and Information Science*. Springer Science and Business Media Deutschland GmbH, pp. 621–632.

Dashtipour, K., Gogate, M., Adeel, A., Larijani, H. & Hussain, A. (2021a). Sentiment analysis of persian movie reviews using deep learning. *Entropy*. MDPI AG, vol. 23(5).

Devlin, J., Chang, M.-W., Lee, K., Google, K. T. & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [online].Available at: https://github.com/tensorflow/tensor2tensor.

Djaballah, K. A., Boukhalfa, K. & Boussaid, O. (2019). Sentiment Analysis of Twitter Messages using Word2vec by Weighted Average. *2019 6th International Conference on Social Networks Analysis, Management and Security, SNAMS 2019*. Institute of Electrical and Electronics Engineers Inc., pp. 223–228.

Hasan, A., Moin, S., Karim, A. & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications*. MDPI AG, vol. 23(1), p. 11.

Ibrahim, M., Bajwa, I. S., Ul-Amin, R. & Kasi, B. (2019). A Neural Network-Inspired Approach for Improved and True Movie Recommendations. *Computational Intelligence and Neuroscience*. Hindawi Limited, vol. 2019.

Kalpana B, Panwar, N., Shalini, R., Rishika V & Pooja K. (2022). Twitter And Instagram Sentiment Analysis of Covid. *Journal of University of Shanghai for Science and Technolog*.

Mhamed, M., Sutcliffe, R., Sun, X., Feng, J., Almekhlafi, E. & Retta, E. A. (2021). Improving Arabic Sentiment Analysis Using CNN-Based Architectures and Text Preprocessing. *Computational Intelligence and Neuroscience*. Hindawi Limited, vol. 2021.

Oueslati, O., Cambria, E., HajHmida, M. ben & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*. Elsevier B.V., vol. 112, pp. 408–430.

Oussous, A., Benjelloun, F. Z., Lahcen, A. A. & Belfkih, S. (2020a). ASA: A framework for Arabic sentiment analysis. *Journal of Information Science*. SAGE Publications Ltd, vol. 46(4), pp. 544–559.

Rehman, A. U., Malik, A. K., Raza, B. & Ali, W. (2019). A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. *Multimedia Tools and Applications*. Springer New York LLC, vol. 78(18), pp. 26597–26613.

Rieser, V. & Refaee, E. (2014). *An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis Related papers Subject ivit y and Sent iment Analysis of Arabic T wit t er Feeds wit h Limit ed Resources Can We Read Emot ions from a Smiley Face? Emot icon-based Dist ant Supervision for Subject ivit y and … Evaluat ing Dist ant Supervision for Subject ivit y* *and Sent iment Analysis on Arabic T wit t er Feeds An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis* [online].Available at: https://dev.twitter.com/.

Safaya, A., Abdullatif, M. & Yuret, D. (2020). *KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media*. Online [online].Available at: https://github.com/nlpaueb/greek-bert.

Tomihira, T., Otsuka, A., Yamashita, A. & Satoh, T. (2020). Multilingual emoji prediction using BERT for sentiment analysis. *International Journal of Web Information Systems*. Emerald Group Holdings Ltd., vol. 16(3), pp. 265–280.

Wolny, W. (2016). SENTIMENT ANALYSIS OF TWITTER DATA USING EMOTICONS AND EMOJI IDEOGRAMS [online].Available at: www.twitter.com.

Zhang, Y., Song, D., Zhang, P., Li, X. & Wang, P. (2019). A quantum-inspired sentiment representation model for twitter sentiment analysis. *Applied Intelligence*. Springer New York LLC, vol. 49(8), pp. 3093–3108.