SENTIMENT ANALYSIS OF ARABS IN TURKEY USING DEEP LEARNING ON SOCIAL MEDIA DATA

2022
MASTER THESIS
COMPUTER ENGINEERING

İnas CUMAOĞLU

Thesis Advisor
Assoc Prof. Dr. Yüksel ÇELİK

SENTIMENT ANALYSIS OF ARABS IN TURKEY USING DEEP LEARNING ON SOCIAL MEDIA DATA

İnas CUMAOĞLU

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis

Thesis Advisor
Assoc. Prof. Dr. Yüksel ÇELİK

KARABUK
September 2022

I certify that, in my opinion, the thesis submitted by İnas CUMAOĞLU titled "SENTIMENT ANALYSIS OF ARABS IN TURKEY USING DEEP LEARNING ON SOCIAL MEDIA DATA " is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Yüksel ÇELİK .....
Thesis Advisor, Department of Computer Engineering

Asst. Prof. Dr. Vedat TÜMEN .....
Second Thesis Advisor, Department of Computer Engineering, Bitlis Eren University

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis September 23, 2022

Examining Committee Members (Institutions)	Signature
Chairman : Assoc. Prof. Dr. Kubilay DEMİR (BEU)	.....
Member : Assoc. Prof. Dr. Yüksel ÇELİK (KBU)	.....
Member : Asst. Prof. Dr. Vedat TÜMEN (BEU)	.....
Member : Asst. Prof. Dr. Kürşat M KARAOĞLAN (KBU)	.....
Member : Asst. Prof. Dr. Omar DAKKAK (KBU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof. Dr. Hasan SOLMAZ .....
Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

İnas CUMAOĞLU

ABSTRACT

M. Sc. Thesis

SENTIMENT ANALYSIS OF ARABS IN TURKEY USING DEEP LEARNING ON SOCIAL MEDIA DATA

İnas CUMAOĞLU

Karabük University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisors:
Assoc. Prof. Dr. Yüksel ÇELİK
Asst. Prof. Dr. Vedat TÜMEN
September 2022, 65 pages

Social Media (SM) has attracted many people from different races and cultures as people consider SM sites as a sharing point where they can freely and uninterruptedly express their opinions. Since many shares are made on SM sites, high-dimensional data is generated. Innovative and more efficient methods for processing this data are being developed. Researchers are interested in sentiment analysis studies that allow more effective and accurate analysis.

This thesis presents a new data set about Arab opinions about Turkey collected from Twitter and the Arabic Sentiment Analysis (ASA) on this dataset. Twitter, one of the most important social networking sites today, allows collecting tweets via API, analyzing them differently, and developing software studies. Our data set is multi-dialectic Arabic and has multiple fields, such as Turkish economy, tourism, food, and

politics, which makes our job even more difficult. After collecting the tweets, they were hand-crafted as positive and negative emotions according to their content and converted into an annotated data set. The data set contains 3136 tweets divided into 1583 positive and 1553 negative ones.

A deep learning-based Arabic sentiment analysis (ASA) approach has been proposed to classify the newly created dataset as positive or negative emotions according to its content. Word2Vec and Bidirectional Encoder Representations (AraBERT) are used in the proposed ASA approach to extract features. Then, bidirectional long short-term memory, Convolutional neural networks, and feedforward neural networks were applied for the binary classification. In addition, a transformer auto classifier based on AraBERT has been applied for the ASA approach. This study determined that the pre-trained AraBERT outperformed Word2Vec and the automatic classifier provided the highest accuracy.

Key Words : Artificial intelligence, Arabic sentiment analysis, data mining, deep learning, Natural language processing, Word embedding, social media, Twitter.

Science Code : 92432

ÖZET

Yüksek Lisans Tezi

SOSYAL MEDYA VERİLERİ ÜZERİNDE DERİN OĞRENME KULLANILARAK TÜRKEYEDEKİ ARABLARIN DUYGU ANALİZİ

İnas CUMAOĞLU

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanları:
Doç. Dr. Yüksel ÇELİK
Dr. Öğr. Üyesi Vedat TÜMEN
Eylül 2022, 65 sayfa

Günümüzde insanlar sosyal medyayı fikirlerini özgürce ve kesintisiz olarak ifade edebilecekleri bir paylaşım noktası olarak değerlendirmektedir. İnsanlar anlık haber takibi, haberlerin ya da yapılan paylaşımların altına duygu ve düşüncelerini yorum olarak yazmaktadırlar. Sosyal Medya sitelerinde çok sayıda paylaşım yapıldığından yüksek boyutlarda veri oluşmaktadır. Bu verilen işlenmesi için yenilikçi ve daha etkin yöntemler geliştirilmektedir. Aaraştırmacılar, daha etkili ve doğru analizlere olanak sağlayan duygu analizi çalışmalarıyla ilgilenmektedir. Bu tez, Türkiye'nin Arap görüşleri hakkında Twitter'dan toplanan yeni bir veri seti ve bu veriseti üzerinde yapılan Arapça duygu analizlerini sunmaktadır (Arabic Sentiment Analysis, ASA). Günümüzde en önemli sosyal paylaşım sitelerinden biri olan Twitter, API üzerinden twitleri toplamasına ve bunlar üzerinden farklı analiz yapılmasına ayrıca yazılım çalışmaları geliştirmelerine olanak sağlamaktadır. Veri setimizin çok lehçeli Arapça

olması ve Türkiye ekonomisi, turizm, gıda ve siyaset gibi çoklu alan olması işimizi daha da zorlaştırmaktadır. Twitler topladıktan sonra içeriğine göre olumlu ve olumsuz duygular şeklinde elle etiketlenerek veri setine dönüştürülmüştür. Veri seti 1583 olumlu ve 1553 olumsuz olmak üzere toplam 3136 tweet içermektedir. Yeni oluşturulan veri setini içeriğine göre olumlu veya olumsuz duygular olarak sınıflandırmak için derin öğrenmeye dayalı bir Arapça duygu analizi (Arabic sentiment analysis, ASA) yaklaşımı önerilmiştir. Önerilen ASA yaklaşımında, öznitelik çıkarmak için Word2Vec ve Arapça Dönüştürücülerden Çift Yönlü Kodlayıcı Temsilleri (AraBERT) kullanılmıştır. Daha sonra ikili sınıflandırma için çift yönlü uzun kısa süreli bellek, Evrişimli sinir ağları ve ileri beslemeli sinir ağları uygulanmıştır. Ayrıca, ASA yaklaşımı için AraBERT tabanlı bir transformatör otomatik sınıflandırıcı uygulanmıştır. Bu çalışmada, önceden eğitilmiş AraBERT'in Word2Vec'ten daha iyi performans gösterdiği ve otomatik sınıflandırıcının en yüksek doğruluğu sağladığı tespit edilmiştir.

Anahtar Kelimeler : Yapay zeka, Arapça duygu analizi, veri madenciliği, derin öğrenme, Doğal dil işleme, Kelime yerleştirme, sosyal medya, Twitter.

Bilim Kodu **:** 92432

ACKNOWLEDGMENT

At the outset, I thank Allah for helping me to complete the research stage, hoping that this work will be of benefit and blessing.

And I would like to express my gratitude to my adviser, Assoc. Prof. Dr. Yüksel ÇELİK and to my secondary supervisor Asst. Prof. Dr. Vedat TÜMEN, for his keen attention and support during the writing of this thesis.

CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVITIONS INDEX

AI	: Artificial Intelligence
SM	: Social Media
ML	: Machine Learning
DL	: Deep Learning
MSA	: Modern Standard Arabic
PP	: Pre-Processing
CBOW	: Continuous Cag Of Words
CNN	: Convolutional Neural Network
RNN	: Recurrent Neural Network
LSTM	: Long-short-Term Memory
BiLSTM	: Bidirectional Long Short-Term Memory
FFNN	: Feed-forward neural network
AraBERT	: Arabic Bidirectional Encoder Representations From Transformers Model
WE	: Word Embedding
SVM	: Support Vector Machine
KNN	: K-Nearest Neighbors
SGD	: Stochastic Gradient Descent
TF-IDF	: Term Frequency Inverse Document Frequency
NB	: Naive Bayes
RF	: Random Forest

PART 1

INTRODUCTION

In recent years, we have seen widespread use of social media (SM) platforms, where people find a haven to express their feelings, personal lives, points of view, and opinions[1].One of the most famous social networking platforms is Twitter, established in 2006 by Jack Dorsey. Year after year, the number of Twitter subscribers increased because it allows the free sharing of ideas and political and religious trends. It is a virtual community that brings people together with artistic[2], political, and sports celebrities through which they can follow the news of those celebrities [3]SM is propelled by Artificial Intelligence (AI). It has become an essential component of daily life, where people face the interference of AI in their daily lives [4], such as querying the weather, using GPS, or searching on the Internet about a topic. In short, AI has become programmed into our lives.

One of the most prominent AI sciences is Natural Language Processing (NLP), which depends on understanding, analyzing, and dealing with human languages using AI and its branches, such as Machine Learning (ML) and Deep Learning (DL) [5]. Whereas ML depends on training the model on big data and then experimenting with evaluating the model's performance, NLP depends on building a model capable of assimilating the language. DL is an extension of neural networks, but it is more profound and complex, capable of learning very complex models, but it needs enormous data.

NLP has many applications: email filters to organize emails and detect spam; smart assistants like Siri and Alexa; text predicting to facilitate and speed up writing and translation. Sentiment Analysis (SA) is our topic and essential research to summarize texts, books, or articles without losing meaning. In order to determine whether a text is expressing positive, negative, or neutral thoughts about a topic, SA integrates NLP, computational linguistics, and textual analysis [6]. In other words, it is the term that

expresses human opinions and emotions, perspectives, attitudes, or ideas. Everyone has a positive or negative opinion of a product, service, or movie. Also, political and religious trends differ from one person to another, so SA, or what some researchers call opinion mining, has received significant attention lately due to the need for many companies to know the evaluation of users of their services or products in order to improve and develop them [7] At the same time, politicians' resort to applications of analysis Feelings in order to predict the results of voting and elections early [8].

Turkey is one of the most important countries that attract tourists from all over the world. Because of the Islamic character and many mosques there, also the Turkish culture, which is very similar to the Arab culture, Arab tourists in Turkey found something that satisfies their desires and longings. Therefore, in this thesis, we turned to analyze the sentiments of the Arabic text collected from Twitter for Arab tweeters. We extracted useful tweets and then applied DL algorithms to categorize the tweet into positive or negative feelings.

1.1. PROBLEM STATEMENT

- The lack of a corpus of the Arabic language
- Most people use their local dialects to express their opinions on SM.
- Poor ability to analyze different Arabic dialects
- The difference in cultures and customs between Arab and Turkish societies can cause problems and misunderstandings between the two societies, especially since Turkey is a country that attracts Arabs from different Arab countries, whether for tourism, study, treatment, or even the reception of refugees.

1.2. THE AIM OF STUDY

- It is building a multi-dialect Arabic corpus that contains Arab feelings toward Turkey.
- Helping the Turkish government understand the Arab opinion on SM and the difficulties faced by the Arab community in Turkey. Moreover, this educates the Turkish people to understand the points of difference between the Arab and

Turkish peoples, thus ensuring safety and compatibility between the two societies, as well as increasing Arab interest in heading to Turkey, which will lead to Turkey's economic prosperity.

- We are building a DL system that analyzes sentiment in many Arabic dialects.

1.3. THESIS ORGANIZATION

The first section of the thesis included an introduction to SA and NLP and the research problem and objective. Then, in the second section, we moved to a review of the newest literature related to SA of Arabic texts. In contrast, the third section included a comprehensive overview of NLP, SA, and its types, as well as the methodologies used in analyzing the feelings of Arabic texts, in addition to presenting the challenges facing the researcher in dealing with SA of the Arabic text. The fourth section included the methodology we followed in analyzing feelings for the new Arabic data set, which we collected and hand-crafted categorized. In the fifth section, we presented the results we reached, which were also discussed in the same section. In the sixth section, we added the conclusion and our aspirations for future work.

PART 2

LITERATURE REVIEW

Thirteen Arabic datasets in Modern Standard Arabic (MSA) and dialect were combined from various sources, including reviews of books, movies, restaurants, hotels, products, and tweets. The new dataset contains 92123 samples. Omara et al. pre-processed the data to remove repeated letters, URLs and HTML tags, stop words, and non-Arabic letters, normalizing and tokenizing. They suggested using three deep Convolutional Neural Networks (CNNs) for Arabic sentiment analysis (ASA) without word level but by character level features. They applied the model to the pre-processed data set and the raw one. Then compared the suggested model with (ML) classifiers: (Nearest Neighbors, Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Multinomial Naïve Bayes, Logistic Regression (LR), and Bernoulli Naïve Bayes). The proposed model shows around 7% accuracy enhancement compared to ML classifiers [9]

Ombabi et al. used a multi-domain Arabic sentiment corpus from studies such as hotel, restaurant, movie, product, and LABR datasets. The dataset contains 15100 entries divided into 7550 positive and 7550 negative scales. To re-processing, they used Arabic NLTK in order to remove stop words and duplicated letters like (كثيييييييرا, رهيييييييييييب), normalization, remove non-Arabic letters and to correct misspelled writing. The authors suggested the Deep (CNN) with Long Short-Term Memory (LSTM) CNN-LSTM Arabic-SA model built as one layer CNN with two LSTM layers to activate local memorization. They used Embedding FastText (Skip-gram and CBOW) as the input layer of CNN. This model's accuracy was 90.75%, which was a good result. The authors tried the effectiveness of the proposed model utilizing other classifiers, Naive Bayes (NB) and K-Nearest Neighbors (KNN), yet SVM was the classifier with the most performing productivity [10]

Bolbol and Maghari compared the performance of three supervised ML classifiers: Division Tree (DT), KNN, and (LR). They applied the classifiers to four Arabic datasets: Arabic Sentiment Tweets Dataset (ASTD), which contains 10K Arabic tweets with objective, positive, negative, and mixed tweets, but the authors applied the classifiers only to the positive and negative tweets as the number of positive tweets is 799 and the number of negative tweets is 1684. Another dataset is Arabic Jordanian General Tweets (AJGT). It has 1800 tweets in the Jordanian dialect divided into 900 positive and 900 negative tweets. Also, they used Arabic Sentiment Twitter Corpus (ASTC), which contains 58000 tweets in the Arabic language, but they chose 22000 positive and 22000 negative tweets. The last dataset is Arabic 100K Reviews, a combination of book, hotel, product, and movie reviews. This dataset has 100K reviews with three labels, but the authors chose only the positive and negative reviews as 33333 for each. First, they pre-processed the data according to steps like tokenization, stemming, and a bag of words, then used Term Frequency Inverse Document Frequency (TF-IDF) to select features. By applying the three classifiers to the mentioned datasets, they found that the LR classifier gave the best accuracy on the Arabic 100K Reviews dataset compared with KNN and DT [11]

By gathering tweets from Twitter expressing concern about COVID-19 using the Twitter API and Python, Al-Sorori et al. sought to analyze Arabic sentiment regarding the virus. They fetched 8046 tweets, filtered them to 770 tweets, then manually annotated the document to divide tweets into positive class when the tweet carried a feeling of anxiety or fear and negative class when the tweet did not have any or neglected the anxiety or fear feeling. For Pre-Processing (PP), they removed stopwords, punctuations, and non-Arabic letters in addition to tokenization and stemming by the NLTK library. They employed CBOW, a Word Embedding (WE) technique, on two pre-trained models. Consequently, they generated the vector of features, and each token is represented in a word vector. They used single, and ensemble classifiers: single classifiers are Nu-Support Vector Classification, linear Support Vector Machine, Logistic Regression cross-validation, Stochastic Gradient Descent (SGD), and Bernoulli NB.On the other hand, ensemble classifiers are (RF) and voting classifiers with hard and soft types. They evaluated the model's performance by 10-fold cross-validation after dividing the dataset into 90% for

training and 10% for testing. At last, the experimental results showed that when applying word2vector with SMOTEEN on CC.AR.300 and in Arabic News, the performance was better than applying WE without SMOTEEN [12]

Alshammari and Almansour presented a new Arabic dataset containing customers' opinions about telecom companies in Saudi Arabia. The data was collected using Twitter API, python, and R language, and they relied on tweets related to customers' feelings toward Saudi telecom companies. They also deleted retweets and repeated tweets and then applied them to PP, such as removing non-Arabic letters and symbols to reduce the feature using the R language, after which they classified the tweets. The data set size was 1096 tweets divided into three classes manually, which are positive, negative, and natural. The third step was the training and testing phase, where they divided the dataset into 80% for training and 20% for testing, as they extracted the feature using Unigram and Bigram. Then through TF-IDF, the training and testing matrix was generated. At last, they used ML classifiers, RF, SVM, LR, and DL algorithms: the neural network, WE, and part of speech. After applying them, they found that DL with WE gives the highest accuracy, with 81% [13]

To represent the word Elfaik and Nfaoui, CBOW and the classified sentiment by Bidirectional long short-term memory (BiLSTM), one LSTM unit for input and another for backward, with an attention mechanism used in the stage of producing the weighted representation to find the relevant part on the sentence. They used three Arabic pre-trained datasets: ASTD, ArTwitter, and Main AHS Arabic Health Services. Furthermore, they compared the suggested approach: (AraVec with biLSTM and attention) with ML classifiers: Gaussian naïve Bayes (GNB), RF, NU-support vector, LR and (SGD), and DL algorithms, which are CNN, RNN, LSTM+CNN, and biLSTM then the experiment showed that the suggested approach gave the best accuracy [14]

To take an interest in improving the Saudi economy, Alharbi and Qamar proposed examining the sensations of clients who visit cafes and eateries in the Qassim district by directing an assessment study through Microsoft form in Arabic. Consequently, a dataset with a size of 1785 surveys was made, and afterward, it was decreased after sifting and pre-handling to 1,507 audits. Using five ML models (KNN, NB, LR, AVM,

and RF) after TF-IDF extract features, they could reach 89% accuracy by SVM, 93% F-measure by RF, and 92 % recall likewise by SVM [15]

Sayed et al. made a dataset from the clients' reviews on the Booking.com site to arrange the text into two classes: positive and negative. Consequently, an Arabic corpus was made in the casual vernacular and MSA with the name (RSAC), a shortened form for Review Sentiment Analysis Corpus, and incorporates 6318 surveys separated into 3354 positive and 2964 negative. After the PP, they utilized TF_IDF with uni-grams, then, at that point, prepared the model utilizing nine models of ML 6 conventional (NB, LR, RF, SVM, KNN, DT) and three new ones not utilized beforehand in text classification or SA. (Ridge classifier (RC), Multi-layer Perception (MCL), Gradient Boosting (GB)). Finally, they examined the performance several times, once without any PP and again with stop words removed, stemming effect, and utilizing all PP approaches, comparing accuracy, recall, precision, and F1- score, and discovered that RC was the best with 95 percent accuracy [16]

The creators of [17] recommended utilizing ML models to know the Sudanese individuals' assessment of the ridesharing service. So, they gathered Arabic data sets from Twitter in their Sudanese vernacular. The dataset size was 2116 in total, divided into 768 positive, 841 negatives, and 507 neutral tweets, and collected from that data set 686 stop words in the Sudanese lingo that were added to the MSA stop words set. Then the pre- Processing stage began, and afterward, TF_IDF with n-gram to extract features. They used KNN, NB, and SVM in the classification stage to categorize tweets as positive, negative, or natural. They tried the proficiency of the models after following a few procedures from PP by estimating accuracy and F1 score. They saw that SVM introduced higher productivity contrasted with the other two classifiers, with an exactness of 95% in the wake of applying stemming as PP.

Alassaf and Qamar suggested using aspect-based sentiment analysis (ABSA) to analyze the opinions and sentiments of Qassim University students. They began collecting data on Al Qassim University from Twitter using the Twitter API, and it grew to 8,234 tweets before being reduced to 7,934 after eliminating tweets whose classification differed between annotators and after the PP phase. The job was divided

into two tasks: the first sub-task was aspect detection Thus chose, nine aspects for examination in this work, including environment, educational aspects, and activities. On the other hand, the second sub-task was aspect-opinion classification, so they defined the polarity of the detected aspect by suggesting that the classes be labeled as negative or non-negative polarity. The few positive tweets were mixed in among the neutral tweets in the non-negative category. They utilized SVM and ANOVA to identify features and determine which of the nine aspects the tweet belonged to. They also utilized the F-score 20 percent Reg approach to reduce features from 161,396 to 7,361. The model's efficiency with SVM was then measured using cross-validation using 10-fold [18]

Using the Bidirectional Encoder Representations From Transformers Model (BERT) model, Das et al could predict Biden's victory over Trump in the presidential election in 2020 by analyzing the sentiment of political tweets [19].

Abdelli et al. used a dataset combined with three datasets available to other researchers with a dataset created by the authors that they collected from Facebook comments. The complete balanced dataset was 49,864 in the Algerian dialect and MSA. They used the data set to classify the text after dividing it by 85% for training and 15% for testing. On the other hand, they used another data set combined from five data sets that were also previously in the size of gigabytes to be used in training CBOW. The authors wanted to compare the performance of SVM and TF-IDF with LSTM and CBOW in ASA. The experiment showed that SVM gave higher accuracy than LSTM accuracy[20]

PART 3

THEORETICAL BACKGROUND

The papers mentioned in the preceding chapter demonstrated how beneficial the findings of this research are for comprehending SA approaches. Moreover, we discovered how SA improves several aspects of life. This section will convey this knowledge and computations and outline the steps to design a SA system.

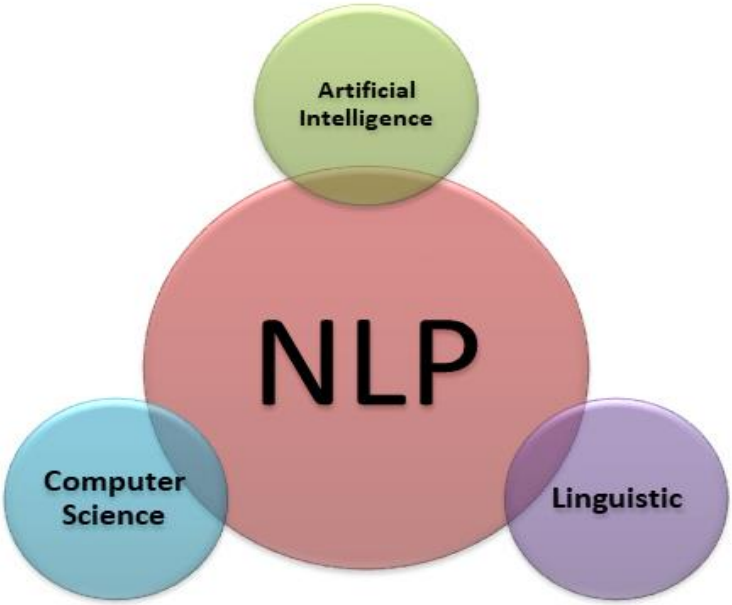

Figure 3.1. The configuration of NLP.

3.1. NLP TYPES

NLP is divided into natural language understanding (NLU): which aims to understand human language, whether text or sound, and interpret it so the machine can understand. Hence, it deals with language as input [20]. In contrast, the second type: natural language generation (NLG), aims to make the machine capable of generating a text or

sound understood by the human being. It deals with language as an output only [21], [22].

3.2. MULTIPLE TECHNOLOGIES THAT USE NLP

- Speech recognition

It converts voice data into text data to save typing time[23].

- Text classification

Helpful in searching for specific parts or keywords within meaningful content or books and texts[24].

- Machine translation

Automatic translation between multiple languages gives the most accurate translation of the source language [25].

- Sentiment Analysis

Understanding feelings through text or speech, therefore, classifying them into positive, negative, or neutral feelings or categorizing them into several categories such as sadness, joy, and anger [26].

- Automatic Text Summarizing

Sometimes we want to understand the general idea of a large text, so shortening and summarizing content or text saves time and effort [27].

- Part of speech

Responsible for dividing the sentence into parts of a verb, a letter, a noun, an adjective, or an adverb to facilitate machine understanding of the written text [28].

3.3. BACKGROUND OF SENTIMENT ANALYSIS

It is one of the modern sciences related to AI and ML; it attracts many researchers because of its importance in several sectors and fields. It provides a comprehensive view of opinions on a particular topic or field. It is considered the sixth sense for those interested in politics or economics. It guides the customer who wants to purchase a product or subscribe to a service. SA performs by using ML and DL algorithms to analyze the texts available on websites and SM. Therefore, to analyze feelings through texts, several steps must be followed to reach the intended goal. Figure 3.2 shows these steps, which we will discuss in detail.

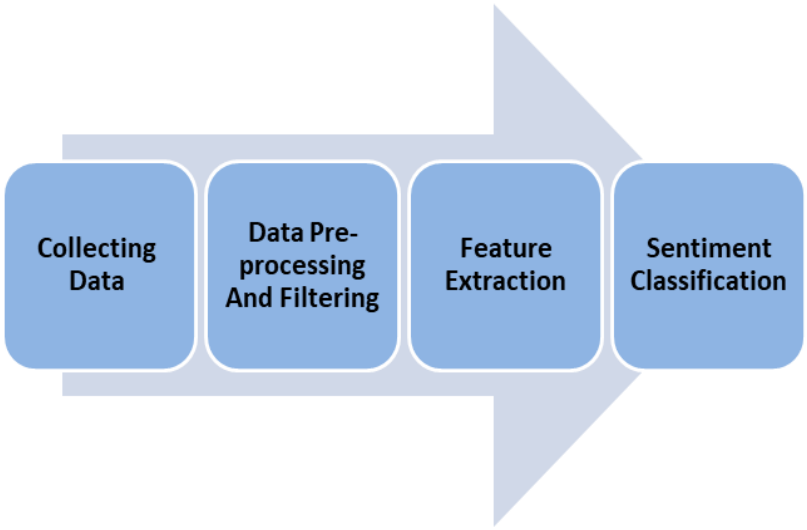

Figure 3.2. SA general steps.

3.4. LEVELS OF SENTIMENT ANALYSIS

SA classification level is included in 3 classes, as shown in figure 3.3, and explained as follows:

- Document Level Classification gives positive or negative classification throughout the whole document. In other words, one opinion of the whole document[29].

- Sentence Level Classification occurs in two steps: firstly, determining whether the sentence is an objective or subjective sentence [30] where an objective sentence contains facts and a subjective sentence has opinions. Then if the sentence is subjective, the polarity will be assessed.
- and Aspect Level Classification at this level, specific entities from several aspects will be processed and determine the opinion according to many aspects [31].

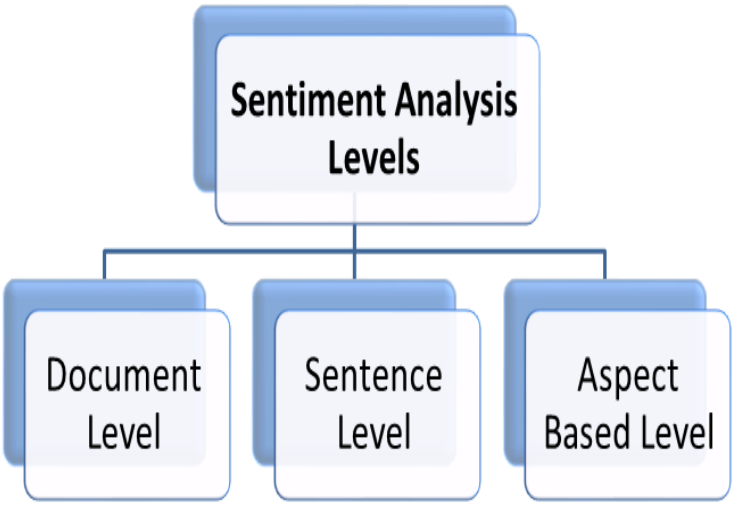

Figure 3.3. SA levels.

3.5. SENTIMENT ANALYSIS AND SOCIAL MEDIA

Notably, Ward Christensen and Randy Suess developed the bulletin board system (BBS) in 1978, which was used to notify friends of gatherings and disseminate information through postings, marking the beginning of social networking platforms. It is the first interactive system. The year 1978 saw the development of the bulletin board system (BBS) by Randy Suess and Ward Christensen, which was used to notify friends of gatherings and disseminate information through postings, marking the beginning of social networking platforms. It is the first interactive system[32]. Then other famous platforms began to appear one by one. Table 3.1 shows the date of the appearance of some platforms and their owners.

Table 3.1. SM sites with their owner and release year.

NAME	DEVELOPER/OWNER	YEAR
SixDegrees.com	Andrew Weinreich	1997
MSN Messenger	Microsoft	1999
LinkedIn	Reid Hoffman	2003
Facebook	Mark Zuckerberg	2004
Twitter	Jack Dorsey	2006
Pinterest	Ben Silbermann	2010
Instagram	Kevin Systrom	2010
Snapchat	Evan Spiegel	2011
Twitch	Justin Kan	2011
TikTok	the Chinese technology company ByteDance	2017

SM is like a notebook where people share their joys, sorrows, successes, and opinions. These platforms are critical as they contain massive amounts of data, but they are not structured, so data analysts organize them and then analyze and classify them. Therefore, we can say that SA applications are closely related to SM platforms. Without SM and its subscribers, the big data used by data analysts would not be available. Thus, large companies can reach their interested customers through advertising campaigns on SM, and the campaigns reach the right people through applications for SA for these people. We also note that SM subscribers have become more open and aware, as when a person wants to buy a product or subscribe to a service, he reviews the evaluations available by people who have subscribed to that service or tried that product before them. So, the profits of those companies increase or decrease according to those evaluations. In other words, the primary concern of modern and advanced companies is to know the customer's opinion of their products and the quality of the service provided. Companies that seek to understand the opinions of their customers understand their needs and constantly strive to make them comfortable and happy. Thus, they believe that the customer is the source of profit and are the companies that top the list of leading companies. SA helps companies' owners improve their services and products in case of a complaint or weaknesses in the service

or product. Therefore, these companies seek to solve these problems and gaps. Moreover, to satisfy their customers and increase their profit, the work[33] presented a system with 87% accuracy to evaluate the satisfaction of the customers using one of the ML algorithms (SVM) to classify the sentiment of tweets into positive or negative and used unigram to extract features.

In the same context, election campaigns are spread on SM, affecting the changing of the opinions of millions of people, where the candidate knows, by analyzing the feelings of the public, what he must present in his electoral campaign and what he must improve in terms of the people's living conditions [19]. It also enables the candidate to see the competitors' electoral campaigns, giving him a competitive power to fill in the gaps he finds in the electoral campaigns of his opponents.

3.6. APPLICATIONS OF SENTIMENT ANALYSIS

SA has entered many aspects of life, significantly influenced expectations, and tipped the scales in many sections. Figure 3.4 illustrates some of the aspects of the SA effect.

3.6.1. Politics

SA is essential in making candidates and political parties aware of the people's attitudes and aspirations. Accordingly, candidates can formulate and improve their electoral campaigns throughout the election period [34]to satisfy voters to increase their chances of winning.

3.6.2. Marketing

Opinion mining helps individuals purchase a product or service from a particular company without relying on outside consultants who charge large amounts of money for each consultation[35]. It also helps companies know the extent of their customer's satisfaction and thus solve problems and gaps in previous services and products. SA can also be used with recommendation systems to increase sales and profits.

3.6.3. Healthcare

SA has made outstanding achievements in the health sector. Patients may share their experiences with a disease, including symptoms, medications, the hospitals they have attended, and recommendations for specific doctors [12]. Similarly, when a new disease spreads, people quickly express their thoughts, experiences, fears, or indifference toward it. This information can be used to gauge the severity of the disease or the rate at which it is spreading, as was the case with Corona disease, which drew numerous researchers to study SA in many aspects and different countries.

3.6.4. Finance

Analysts produce stock market predictions in real life based on public opinion and news events[36]. Similarly, the SA applications allow machines to perform the same task. Furthermore, using advanced computational linguistics and ML approaches, opinion mining is more efficient than human analysts, with the ability to scan through large amounts of text across multiple news outlets in seconds.

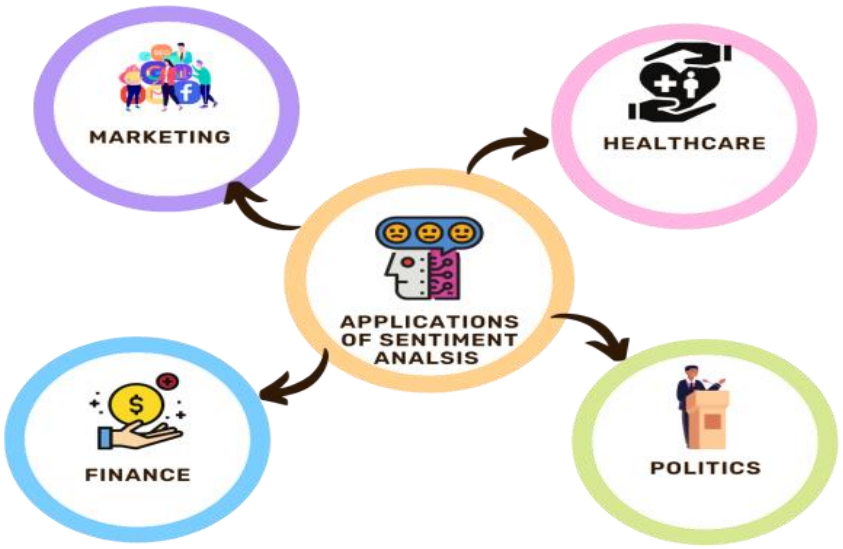

Figure 3.4.Effecting of SA on life aspects.

3.7. SENTIMENT ANALYSIS APPROACHES

3.7.1. Lexicon-Based Approach

This approach is based on collecting many words that carry an emotion or feelings of positive, negative, or perhaps neutral polarity. Therefore, early works have tried to create dictionaries of feelings that relate each term to at least one tendency and sensation[36]. This type of technology has the benefit of not requiring training data.

3.7.2. Machine Learning Approaches

ML is a branch of AI. It is the ability to make a ML without human guidance so that it learns by watching the human's previous experience of the required task by giving it large samples for the required task. It then enters specific data and commands so that the machine can learn and recognize different objects and distinguish between them; it also gives other samples that do not belong to the required task. Hence, it develops itself after analyzing that task mathematically. When a specific problem that has never been faced before arises, the machine does not immediately stop but instead searches for an alternate solution [37]. So, the machine is trained to perform the duties on its own and obtains the ability to address future challenges.ML algorithms can learn independently without defining the rules manually, but rather by learning from past data until they predict their own [38].

3.7.3. Machine Learning Types

ML science has three types: supervised, unsupervised, and semi-supervised.

- **Supervised Learning:** when the task is driven, it depends on training the data by knowing inputs and their labeled outputs [39]. In this type, there are two major tasks: regression and classification.
  - Regression aims to predict the closest value to the actual value according to specific inputs or features. In other words, output has continuous value, like

predicting house prices after giving information such as the number of rooms, size, and location [40].

- Classification divides the data after training into two or more given classes according to specific features, classifying patients according to many features [39] such as temperature, runny nose, and headaches into infected with influenza or not infected.

- **Unsupervised Learning:** occurs when the machine does not have any information about data and aims to cluster data into similar groups. By entering unlabeled data, the machine tries to find similarities and differences between entered data and then clustering data into groups [41].. It means unsupervised Learning is like classification with a fundamental difference. In unsupervised Learning, the machine does not have any information about the output, like the ability of the machine to determine that this news is sports news without previous training and the suggestion of friends on Facebook or suggesting films.
- **Semi-Supervised Learning**: this is a blend of the earlier types [42], as a result, during training, it employs a small amount of labeled data, similar to supervised Learning, and a large amount of unlabeled data, similar to unsupervised Learning [42,43]. This type considerably aids the systems in improving learning accuracy.
- **Reinforcement learning:** In this type, the machine or the agent can discover its environment to provide the best solution for the given task; where this method depends on the idea of reward and punishment [44]; thus, the machine takes steps and implements them, and through the reaction that comes from the surrounding environment, it can determine whether it was wrong or correct and works on developing itself. Like giving an agent a positive bonus for winning a game or a negative bonus for losing

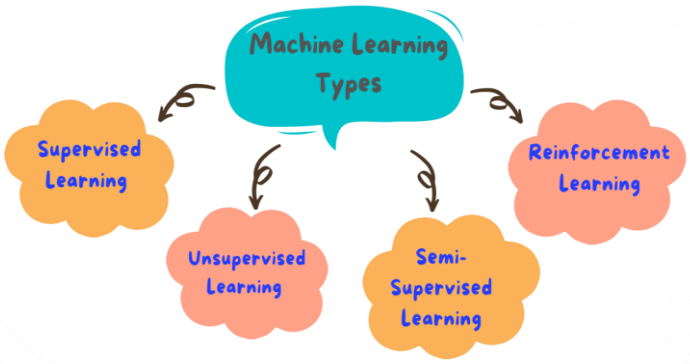

Figure 3.5.ML types.

3.7.4. Deep Learning Approaches

DL is inspired by human neurons and their intertwining with each other. It is a science that falls under the science of ML. This approach uses artificial neural networks consisting of an input layer and an output layer and, between them, the hidden layer or layers that contain the neurons.[45]. DL needs a high volume of data, but its techniques give more accurate results than the standard ML techniques. It proved a high efficiency in computer vision and accuracy in word processing. Recurrent Neural Networks (RNNs) and CNN are the most famous and influential DL algorithms.

3.8. CHALLENGES OF SENTIMENT ANALYSIS

Spam: With the increase in the use of Arabs on SM, Arab content has become popular and contains both legal and illegal content, and therefore we can notice much annoying spam content. Consequently, a tweet on Twitter or a post on Facebook may contain a spam link or a false job advertisement for a fake company that promises its members high salaries and incentives. Despite research in the field of spam detection in the Arabic language is still few and needs to be developed, we may mention the research provided by [46] to detect spam in tweets written in the Gulf dialect, using NB and SVM with more accuracy using NB.

Sentiment polarity fuzziness: The inability to extract feelings or emotion because of the ambiguity of the sentence or containing feelings with opposite polarities [47], such as using the word polarity positive and putting a negative emoji after it or using the

word polarity negative and then a positive emoji like the following sentence that has positive polarity in term sweet then has negative polarity when used crying emoji "الفستان كتير حلو بدي منو 😭" in English "It is so sweet dress I want one like it 😭"

The polarity may be due to the different expressions from one country to another; for example, الله يعطيك العافية means "God give you wellness" in English, which means a prayer of goodness in the Levant, which is a positive feeling. In contrast, it means going to hell in the Maghreb and Libya, feeling negative polarity.

Quality of reviews: Reviews may contain widespread expressions in SM that may have a positive polarity, but categorizing them as such will cause an imbalance in the analysis of the results, for example, تصبحون على خير good night or صباح الخير good morning. They are positive expressions, but then there can be complaints or grumbling, such as "صباح الخير، جربت منتجكم بس ما عجبني سيء كتير" in English "Good morning, I tried your product, but I did not like it very badly."[48] used the SVM algorithm to rank reviews according to their helpfulness.

Sarcasm: It can be considered a type of criticism, a behavior harmful to people's feelings. People's opinions differ; some express them respectfully, unobtrusively, and annoyingly. It is noticeable on SM that there is a large spread of people who love bullying. Irony or sarcasm is a way of expressing an opinion about a product, a political, artistic, or sporting topic, or even an opinion about someone in a crude and tactless manner, using phrases and terms that may sound positive but carry an implied negative emotion. As the following sentence: " شفتي شو كانت لابسة عنجد احلى من هيك ما عاد في هههههه" in English "Have you seen what she was wearing more beautiful than this no longer in hahaha" [49].

Sarcasm is one of the most critical challenges facing SA because it is difficult for the machine to understand what the writer intended behind the words and meanings he wrote, perhaps the opposite of what he wrote. Moreover, detecting irony in Arabic is more complex than detecting it in English because of the complexity of the Arabic language terms and many dialects [50] introduced a new approach to detecting irony in the Arabic language, so they collected tweets and dealt with the problem as a

regression problem, not a classification problem, meaning that they revealed the amount of sarcasm (the level of sarcasm), not its presence or absence.

Implicit Sentiment: When the sentence does not contain a word that carries an explicit emotion, here, the sentiment can be implicit in that sentence, and the sentence can be called an objective[51]. When the sentence does not contain a word that carries an explicit emotion, here, the sentiment can be implicit in that sentence, and the sentence can be called an objective sentence.

Spelling errors: We may notice on social networking sites that there are writing errors because these sites allow the user to write in the way he wants without focusing on the quality of sentences and writing[52] Typos or fast typing can create misspellings because it is located near the original letter on the keyboard; it is treated as a substitution for that letter [53]. In the phrase "Don't be Kate," the letter "l" in the word "late" has mistakenly been changed to the letter "k" to form the phrase. Alternatively, substituting a letter for another because it makes the same sound. For example, "kat" rather than "cat." Additionally, the author can have deliberately misspelled something to accentuate a particular emotion by repeating a letter from a word. For example, "I love you soooooo much" His intense love is indicated by the letter "o" being repeated.
Slangs: It is an informal way of writing as it is used on the web a lot in order to shorten the time or to reduce the number of characters [51] due to the limited number of characters on some sites such as Twitter, which allows the use of a maximum of 280 characters within one tweet; thus slang has become widespread in SM and it is an abbreviation of letters of words[54, 55], for example, "شو عم تعمل" in English "What are you doing" written as "ش ع ت" Which shortened a sentence of three words with three letters in Arabic, Another example of using slang in the English language "Oh my God" to be written as "o m g"

Domain dependency: The polarity of feelings may differ from one domain to another. When using an emotion classifier that has been trained on a specific domain in another domain, it may give wrong results. For example, the word "long " has a positive polarity when describing a young person's adjectives, while it has a negative polarity when used to describe a road [56]. Many studies have used ML approaches to optimize

general-purpose lexicons for a specific domain [57]. Due to the fuzziness introduced by dialects, creating a general-purpose vocabulary for the Arabic language would yield sub-optimal results. It would be more challenging to concentrate on domain and dialects simultaneously.

3.9. ARABIC LANGUAGE AND ARABIC-SPECIFIC CHALLENGES

Arabic is one of the oldest languages in the world, with an estimated age of more than 1500 years. Twenty-two countries speak Arabic as an official language [58], and the number of Arabic speakers is estimated at 420 million; around the world, One of the United Nations' official languages is Arabic [58].

The importance of the Arabic language comes as it is the language of the Qur'an, which is the classical Arabic language; there are 28 letters in the Arabic language, including three vowels. [59]. The word is written in the Arabic language in the form of letters connected but without existing of upper and lower case, and each letter has several ways in written according to its existence in the word, which makes processing the Arabic language more complex than in other languages. Example: ت letter has a تـ shape at the beginning of the word as in تمر, ـتـ shape in the middle of the word as in كتب, also ـت shape at the end of the word as in بيت and ت shape after "ا" letter as in بنات. It is also necessary to mention the presence of diacritics in the Arabic language, which are symbols that are placed above or below the letters to facilitate reading the word correctly and thus facilitating understanding of the meaning. However, the original Arabs do not need the presence of diacritics on words to understand the word [46] because it is their mother tongue. Additionally, the meaning is simple for them to comprehend and know the correct pronunciation without diacritics, so using diacritics in SM is little.

Moreover, some words have the same writing, but their meaning and pronunciation vary according to the diacritics. For example, the word "الجد" has three meanings according to the diacritics of the letter Jim "ج." It means grandfather; if the diacritic on the Jim is fatha "aljad," it means the coast of the sea; if the Jim diacritics is damma "aljod," and it means diligence if the Jim diacritics is kasra" aljid" [60]. However, the

use of the classical Arabic language is very little and is limited to religious and ancient books. The Arabic language adopted in all Arab countries, officially spoken in meetings, conferences, on television, and on radio, is called MSA. While in daily life, Arabic speakers use dialects that differ from one country to another, or they can even differ in the same country from city to city. Accordingly, the dialects in the Arabic language can be divided according to geographical location into five dialects Egyptian Arabic in Egypt and Sudan; Levantine Arabic in Palestine, Jordan, Lebanon, and Syria; Peninsular Gulf Arabic in Saudi Arabia, The United Arab Emirates, Oman, and Kuwait, Iraqi in Iraq and Maghrebi in Morocco, Algeria, Libya and Tunisia [61].

Morphology is the science concerned with analyzing the structure of the word and how it is formed and expressed [64]. We can also call it the science of change because it studies the changes that occur in the word, so it changes its meaning or its syntax. Arabic is considered one of the most complex languages in its construction and rich in morphology [62], so native Arabic speakers may be unable to confront this science because it needs specialists in Arabic language, morphology, and grammar. In general, sentences in the Arabic language are divided into nominal sentence, which is the sentence that contains only nouns without the presence of a verb or a verbal s, and one of its conditions is to contain a verb that distinguishes the sentence. Moreover, Arabic words and verbs can exist in more than one form by deriving from the root of the word (usually three letters) and then using suffixes or prefixes. From the fundamental root of the word, we can get a new meaning or a verb with a new conjugation [62], [63].

The lack of an Arabic language Corpus: A corpus is a huge collection of text from several sources, with information attached to it or any of its constituents such as words, sentences, phrases, documents, and so on [64]. Extracting emotions from texts is deeply dependent on the corpus. However, the Arabic corpus is very few, insufficient, and not comprehensive. Until now, the corpus has not included all types of Arabic: classical, MSA, and colloquial [65].

Using colloquial: Arabs use their dialects to express their feelings, opinions, and points of view, and this makes the task of SA more challenging and complex because these dialects are not subject to standard rules as in MSA [66]. In addition, ordinary people may use strange terms that differ from one environment to another. Perhaps

some people mix several languages in their writings because of the presence of foreign words in the Arabic language that originates from other languages such as English, French, Turkish and Persian, which makes the matter more complicated.

3.10. ARABIC LANGUAGE TEXT PRE-PROCESSING TECHNIQUES

PP eliminates noise in the data after collecting it, it occurs in many steps, and each step has its importance in making the data carry only the important information that will be entered into the model. Thus, at this stage, the data is cleaned of duplicate information, especially in the data collected from Twitter, so they must be removed. Also, links and hashtags downloaded with tweets are removed because they do not carry any importance or feelings. We also note that some people use punctuation or emoji in their writing. Also, in the tweets, the presence of RT, an abbreviation for retweet, and the presence of cc refer to a person [67], is information that is not important in the SA process, so it is removed. Also, the data must be cleaned of stop words, which are common words used frequently. Their presence or absence does not affect the meaning of the sentence. They are words that have no association with feelings, so deleting them will not affect the SA but will speed up the process. It reduces the index space examples of stop words in "من, إلى, كل, نحن, أنا" Arabic and similar to their meaning in English "I, we, every, to, from."

Similarly, some people want to emphasize an emotion, repeat the letter to emphasize the emotion like حلاااااااااااوة, or lengthen the letter because the Arabic language supports several ways to write letters. Thus some want to draw attention by lengthening the letters, or it called Strip elongation (Tatweel) like رهيبـــــــــــة. All of these need to be treated and removed to obtain clean and suitable data to deal with and obtain satisfactory results. Then there are some extra essential steps after cleaning data that are as follows:

Tokenization: is a technique for decomposing a sentence into words, symbols, phrases, and characters. In other words, it divides the sentence into several words (tokens) so that each word becomes separately in the list to facilitate dealing with it later. It is divided into two types: Word Tokenizer, which divides the text into separate words,

and Sentence Tokenizer, which divides the text into separate sentences. Tokenization is similar to the split function but more accurate than split. Consequently, "we're" with split, the output is we're, but when using tokenization, the output is [we, 're]. Figure 3.6 shows an example of using tokenization by spacy and nltk libraries and using the split function.

Spelling check: Many spelling errors, such as repeated letters, can be found in SM data. Therefore, correcting spelling errors is a required task before performing language processing [68].

Stemming: It is the tool that allows the word to be stripped of its affixes whether it is at the beginning or end of a word [69]., therefore returned to its source (root) in the Arabic language, the root of any word consisting of three or four letters.

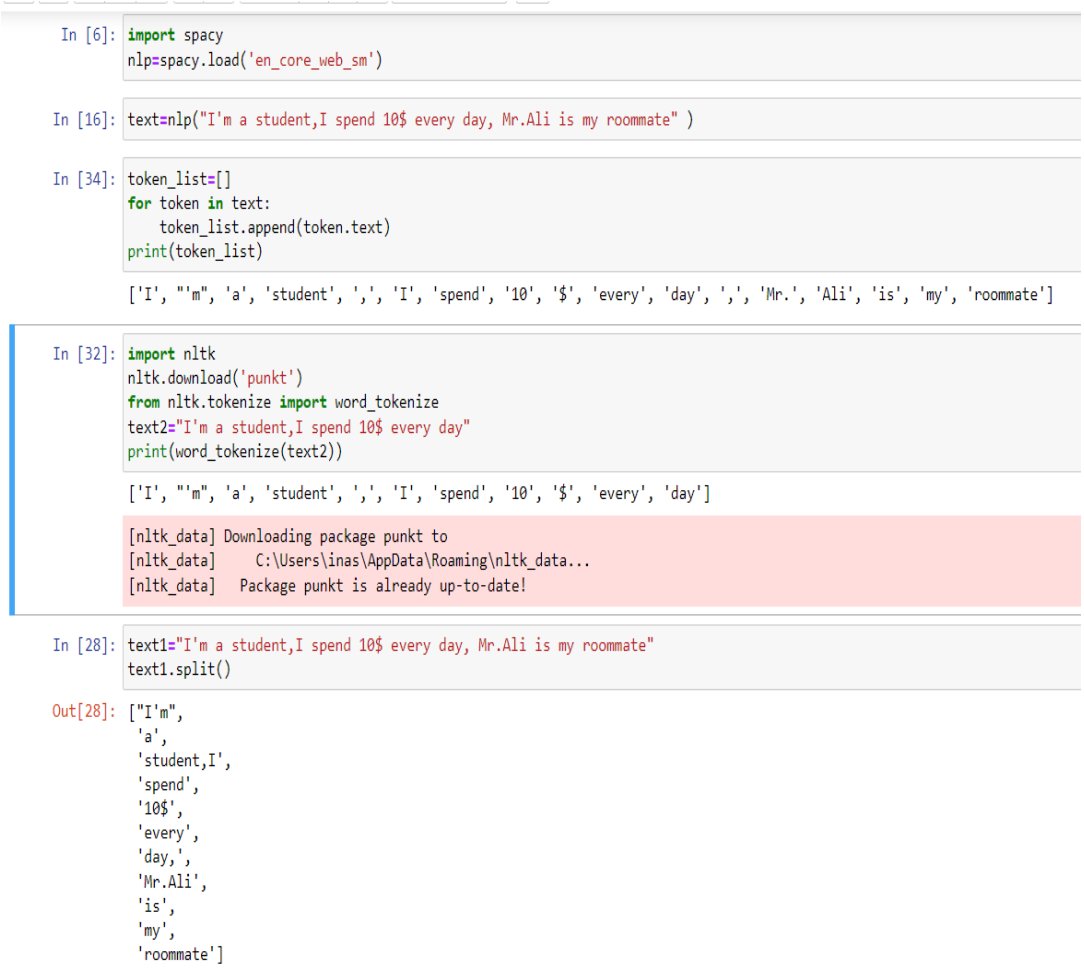

Figure 3.6. Screenshot difference between Token and split.

3.11. TEXT FEATURE EXTRACTION

3.11.1. N-Gram

It is a technique for dealing with words in sentences or with letters in words. Where N denotes numbers and grams denote words/letters. Consequently, where N = 1 is referred to as a unigram, N = 2 as a bigram, and N = 3 as a trigram. Etc table 3.2 shows an example:

Table 3.2. N-gram explanation.

I love learning python so much						
Unigram	“I”	“love”	“learning”	“python ”	“so”	“much ”
Bigram	“I love”	“love learning ”	“learning python”	“python so”	“so much”	
Trigram	“I love learning ”	“love learning python”	“learning python so”	“python so much”		

This technique relies on the science of probability. The probability of each word being repeated after the next word is calculated [70]. Where the N-grams are used in many fields that deal with the production of words or letters, such as:

- Text generations
- Questions answering
- Chabot
- Autocorrect
- Translation

3.11.2. Parts Of Speech Tagging

It is the task responsible for determining the type of word grammatically, a verb, a noun, or an adjective, based on the context of the word in the sentence [71]. It is based on the fact that the meaning of any word is not in itself. But according to its content, context, and the words around it.

3.11.3. Bag Of Words

This technique was used a lot in the beginnings of NLP, but it was less used in modern studies because of the lack of focus on the word's meanings[72] and the slow implementation. Despite the disadvantages of this technology, CBOW is based on it.

3.11.4. Named Entity Recognition

Responsible for identifying and classifying important words such as the names of people or institutions or the names of countries and cities, money, and all famous things and people [73]. It is very used in Wikipedia, as it links famous names with links that take us to them by simply clicking on them or working to link specific questions with answers to them in other links.

3.11.5. Word Embedding

Since ML could not treat texts as words or letters, it was necessary to convert the words into a language understood by the machine, which is the language of numbers, so that the words are converted into vectors in a continuous high-dimensional space [74]. WE is a text-handling technique where convergent vectors represent words with similar meanings by numerical representation (vector representation). It is used to measure the word's semantic meaning and grammatical meaning. Similar vectors by numerical representation represent words with similar semantic meanings. This process uses neural networks to know the words' features [75]. It is worth mentioning that WE have many techniques, as follows:

3.11.5.1. Word2Vec

It is a shallow neural network trained based on WE and is an unsupervised learning method. The input is a text corpus, and the output is several arrays specific to text features. The main task of the word2vec tool is the grouping of matrices of similar, identical, and related words, which is done through the mathematical similarities of each word[76]. Moreover, it aims to calculate the importance and value of each word in the sentence and then infer the remaining word. It also learns whether the word is singular or plural and makes it easier later to make a complete formulation of texts. Along these lines, the semantic closeness of the words to one another is likewise uncovered. The semantic similarity between words is mathematically expressed, for example: "waiter – man + woman= waitress" It is understandable that the terms "waiter" and "waitress" are quite similar. However, their vectorial variances are due to their gender. Two primary techniques in word2vec are the Continuous Bag Of Words (CBOW) and Skip-gram, where CBOW is based on a combination of BOW and N-gram. The main task is to generate a single word after determining the window size that specifies the number of words before and after the required word[77]. For example, if the window size equals 2, two words are given before the required word, and two are given after the required word, as shown in Figure 3.7. It is worth mentioning that the input words will be given as WE array.

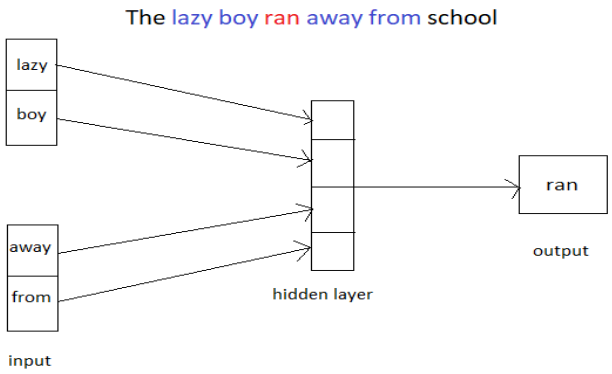

Figure 3.7. CBOW with 2 window size.

On the other hand, Skip-gram works in contrast to CBOW, which generates many words after giving a single input. Here the window size points to the number of words before and after the given single word.

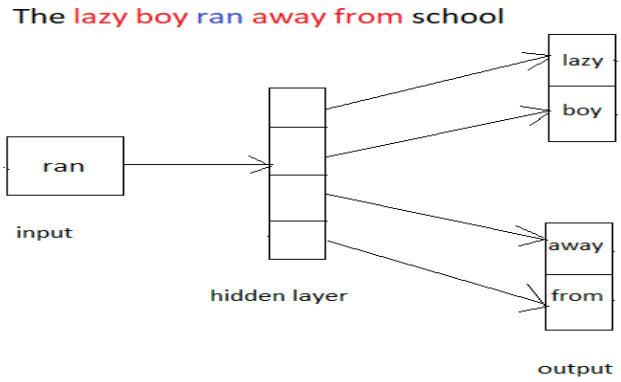

Figure 3.8. Skip-gram with 2 window size.

3.11.5.2. FastText

FastText is a library for professional Learning of word representations and sentence classification. FastText permits train supervised and unsupervised representations of words as well as sentences. These (embeddings) can be utilized for various applications, from information pressure and candidate selection. FastText supports (CBOW) or Skip-gram models. Sub-word generation [78]. We will produce character n-grams of length 3→6 presented in the word. After taking a word, we will add angular brackets to indicate the beginning and end of the word.

sit → <sit>

We will then create character n-grams of length N. By sliding a window of three characters (N=3) from the start of the angular bracket till the closing angular bracket on the word "sitting.

<sitting>
So, as we can see:

3-grams <si sit itt tti tin ing ng>
Sitting

Note: if the word "sit" is a part of the vocabulary, it will be represented as <sit>. This helps safeguard the meaning of short words that might appear as N-grams of different words. Innately, this also lets us know the meaning of postfixes/prefixes.
Different length character n-grams:

Table 3.3. N-gram character level example.

word	N-Length	N-gram
sitting	3	<si ,sit, itt, tti, tin, ing, ng>
sitting	4	<sit, sitt, itti, ttin, ting,ing>
sitting	5	<sitt, sitti, ittin, tting,ting>

As could be a large number of unique n-grams for a word in the text, we should apply to hash; therefore, instead of learning for each n-gram, we learn the whole b embeddings [79].

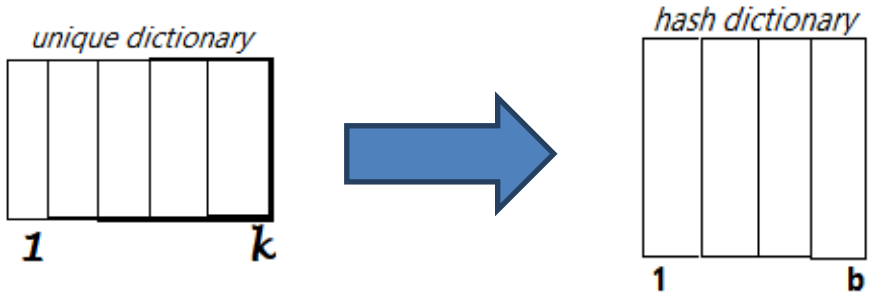

Figure 3.9. Hashed dictionary.

As we can see, every character of n-gram is hashed to an integer that will be between 1 to b.

3.11.5.3. GloVe

It is an unsupervised learning method, an abbreviation for Global vectors of the word representation. It appeared at Stanford University as an open-source project based on the idea of word representation as a vector to discover the similarities between words in context and semantic. The idea of GloVe pre-trained is based on the Co-Occurrence matrix and ratios probability [80]. The Co-Occurrence matrix is square (the rows and columns are equal), consisting of all the words in the text after deleting the repetition. The common values in the table show the extent of their presence with other words in a specific context or a particular sentence.

To calculate the co-existence matrix, we must first determine the size of the window, which is the number of words that will be looked at on the right and left to deal with it. If it is 1, that means we will calculate the relationship of each word with the word before and the next to it, and if it is 2, it means we will calculate the relationship of each word with two words before and two words next to that word, and so on, but provided that they are in the same sentence, table 3.4 and 3.5 show examples of co-occurrence matrix. On the other hand the conditional probability ratios represents in a table 3.6 where the rows represent specific topics and columns represent words and the common value between columns and rows represents the probability of the existence the word in the topic, while the last row represents dividing the first row by the second one, if the result is close to 1 that means the word is return to the both topic or not related to the both topic, and if the value is greater than 1 that means the word is related to the first topic, where the value is too small that means the word refers to the second topic as shown in the example in the table 3.4 where we can notice that win word is related to both sport and political so the result was close to 1, in the same way the word shopping is not related neither to sport nor to political, and football is related just to the second topic (sport) so the result was close to 0, also president usually related to political that is the first topic so the result was huge 80 . In short, GloVe is similar to the WE technique, but it has been trained on a huge number of words [81] to measure the semantical similarity of words to each other.

Table 3.4. Co-Occurrence matrix with 1 window size.

I love learning machine learning and love learning deep learning						
	I	love	learning	machine	and	deep
I	0	1	0	0	0	0
love	1	0	2	0	1	0
learning	0	2	0	2	1	2
machine	0	0	2	0	0	0
and	0	1	1	0	0	0
deep	0	0	2	0	0	0

Table 3.5. Co-Occurrence matrix with 2 window size.

I love learning machine learning and love learning deep learning						
	I	love	learning	machine	and	deep
I	0	1	1	0	0	0
love	1	0	3	1	1	1
learning	1	2	3	2	2	2
machine	0	1	2	0	1	0
and	0	1	2	1	0	0
deep	0	1	2	0	0	0

Table 3.6. Conditional probability ratios.

wk	win	football	shopping	president
p1=P(w_k | political)	0.7	0.005	0.0005	0.4
p2=P(w_k | sport)	0.8	0.8	0.0006	0.005
p1/p2	0.875	0.00625	0.83	80

3.11.5.4. Arabic Bidirectional Encoder Representations from Transformers

Arabic Bidirectional Encoder Representations From Transformers Model (AraBERT) is a BERT-based, pretrained Arabic language model. The algorithm was trained on 3 billion words and 70 million phrases, or 23 GB, of Arabic literature. There are many

versions of AraBERT: AraBERTv1-base, AraBERTv0.1-base, AraBERTv2-large, AraBERTv2-base, AraBERTv0.2-large, and AraBERTv0.2-base.

BERT is based on the transformer. A transformer is a DL model that uses the self-attention process and weights the significance of each component of the incoming data differently. Similar to human attention, neural networks also exhibit some of it. It implies that some inputs are given more attention than others while they leave others [82]. BERT was built based on pre-trained data, which can describe as a deep representation of unlabeled texts, including Wikipedia and book corpus. The model was trained on this large text corpus and is a more profound and good understanding of how the language goes on [83]. Figure 3.10 shows the structure of transformer.

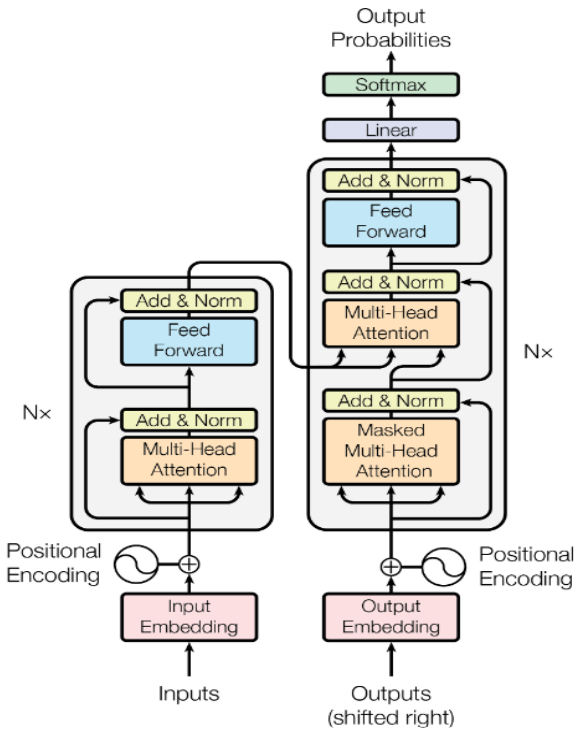

Figure 3.10. The internal structure of transformer [82].

We can describe BERT as a deeply bidirectional model, which means BERT can learn all information from the left and the proper context while training [84]. This feature is significant for a high understanding of the meaning of the text and the language. The BERT model can pick up both the right and the left context of two sentences and then the right understanding of the meaning in each one [85].

There are three embedding layers in AraBERT:

Position Embeddings: Positional embeddings are learned and used by BERT. to specify the placement of phrases in a sentence. These are brought to triumph over the issue of transformer [85].

Segment Embeddings: AraBERT takes the inputs as pairs for tasks like Question-Answering systems, and that helps the model distinguish between these sentences.

Token Embeddings: This Layer can learn for a particular token from the vocabulary of word piece token.

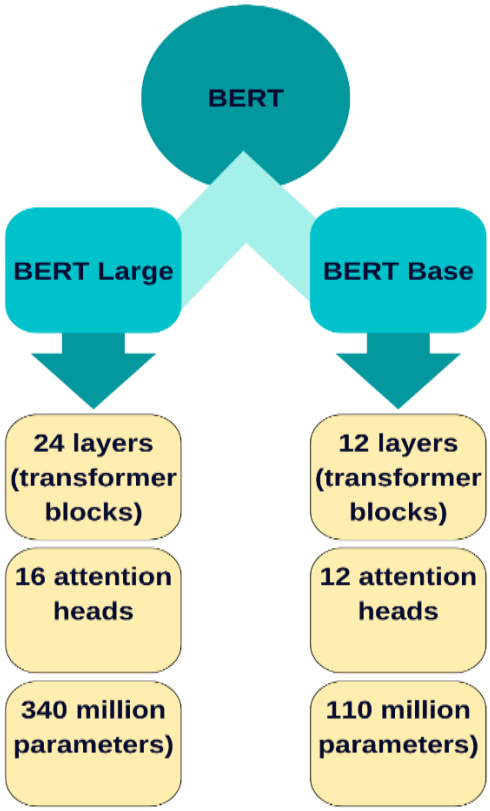

Figure 3.11. The structure of the basic types of BERT.

3.11.6. Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) It is a comprehensive development of the idea of BOW, as it depends on the extent of certain words in the

texts, but more professionally so that it contains the number of times the word is present. It is divided into two methods, term frequency (TF) and inverse document frequency (IDF). TF: is the number of times the word is repeated in the document (Tw) divided by the total number of words in the document (Td).

$$TFw,d = \frac{Tw}{Td} \quad (3.1)$$

Where IDF indicates how rare or common the term (word) is. The word itself, if it is widespread among all texts and its use increases in all other documents, then it has a weak effect on itself. For example, words (لا, نعم, جدا) (very, yes, no,..) are common words with a general meaning and will not affect determining text type. While rare words such as (investment, benefits, bacteria, programming, and grease) are found only in a limited number of documents [86], they have more actual meaning. Words with more frequency in files may have a more significant impact, but in the case that this word is rarely used

$$\text{IDF} = \log\ (\frac{N}{\text{dfw}}) \quad (3.2)$$

N: The total number of documents

df_w: The number of documents that contain the term

At last:

$$TF - IDF = TF \times IDF \quad (3.3)$$

For example, if we have 2000 documents and 20 of them contain a specific word that appears four times in one of the documents that contain 40 words in total, therefore:

TF= 4/40= 0.1

IDF= log (2000/20) = 2

TF-IDF= 0.1*2= 0.2

3.12. MACHINE LEARNING TECHNIQUES FOR ARABIC SENTIMENT ANALYSIS

The most used ML classifiers for ASA are shown in this section.

3.12.1. Naïve Bayes Classifiers

Based on the Bayes theory created by Thomas Bayes (1702-1761), it depends on statistical and probabilistic sciences. This classifier is considered fast, reliable, low cost, and accurate and is a supervised learning algorithm the other [87]. The Naïve Bayes algorithm is based on the fact that the features are unrelated, as each feature is independent of the other. For example, in predicting cancer, the features are the patient's age, gender, smoking, and medical history. Using the NB, these features are not dependent on each other. NB classifier effectively classifies large data and is used in SA, text classification, spam email classification, and recommendation systems. Bayes Formula is as follows:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (3.4)$$

P(A): the probability of event A is true (regardless of the data). The prior probability of h (priori) is what we call it

P(B): the probability of the event B (regardless of the hypothesis). This is known as the prior probability.

P(A|B): the probability of event A given event B. This is known as posterior probability (posteriori)

P(B|A): the probability of event B given that event A was true. This is known as posterior probability.

- Multinomial Naïve Bayes

It uses a multinomial distribution to tackle document or text categorization problems like SA [88].

It is predicated on the idea that the probability of a particular class existing based on a given text will be equal to the probability of the text existing based on the provided class, multiplied by the probability of this class existing based on the text, divided by the probability of the text.

For document d and class c:

$$P(c|d) = \frac{P(d|c).P(c)}{P(d)} \tag{3.5}$$

$P(d|c)$: The probability of the text based on the given class is equal to the probability of the words in the text, based on the class.

$P(c)$: The probability of the class is equal to the number of texts (tweets, comments...) that contain this class, divided by all the number of texts (all tweets, all comments...). $P(d)$:The probability of the text (emotion word) is the number of times that text (emotion word) is repeated over the total number of words in the document.

And the left side $P(c|d)$: is the possibility of a specific class based on the given text (emotion word), such as whether the comment is negative or positive, which is what is required.

$$\begin{aligned} C_{map} &= \arg_{c\in C}\max \hat{P}(c|d) \\ &= \arg_{c\in C} \max P(x_1, x_2, x_3, \dots., x_n|c)P(c) \\ &= \arg_{c\in C} \max \hat{P}(cj)\prod_{x\in X} \hat{P}(x|c) \end{aligned} \tag{3.6}$$

And this equation will be repeated for all classes of the same text (file) to choose the highest probability.

3.12.2. Support Vector Machine

SVM categorizes data by mapping it to a high-dimensional feature space, even when the data is otherwise not linearly separable. The data are processed in a way to let the

separator be represented as a hyperplane [89]. Then a separator between the categories is found. Following that, new data features determine which category new data should be placed in. SVM is the ML approach that successfully categorizes data the most, so we will look at that:

SVM is a supervised learning method. SVM uses an N-dimensional hyperplane to divide data into two categories. The discriminating function in SVM is g(x),

$$g(x) = w^T f(x) + b \tag{3.7}$$

Where w is the weights vector, b is the bias, and f(x) is a nonlinear mapping from the input space to the high-dimensional feature space. The parameters w and b are automatically learned from the training dataset using maximizing margin by minimum.

$$\min \frac{1}{2} W^T W + C \sum_{i=1}^{N} ci \ (2) \tag{3.8}$$

Where C stands for the penalty coefficient and N represents the slack variables. Moreover, because the problem of text classification involves an extensive feature space, the classification issue is always linearly separable; linear kernels are frequently employed. The extraordinary learning ability of SVM can be independent of the feature space's dimensionality. Instead of counting the number of features, SVM assesses the difficulty of the hypotheses based on the margin separating the plane.

3.12.3. Logistic Regression

One of the supervised learning algorithms used for discrete problems and effective in binary predictions like in SA as positive and negative prediction, LR is used to predict the probability of an event occurring with additional knowledge of variable values that can be explained or related to that event [90]. The cost function is defined as follows:

$$\begin{aligned} \text{Cost}\,(h_\theta(x), y) &= -\log\big(h_\theta(x)\big) \quad \text{if } y = 1 \\ &= -\log\big(1 - h_\theta(x)\big) \text{ if } y = 0 \end{aligned} \tag{3.9}$$

The cost function of Logistic Regression is:

$$0 < \mathrm{h}\theta(X) = \frac{1}{1 + \mathrm{e}^{-(\beta 0 + \beta 1 \mathrm{X})}} < 1 \qquad (3.10)$$

3.12.4. Stochastic Gradient Descent

SGD is an effective discriminative learning technique for linear classifiers. The objective function is optimized using an iterative technique. It approximates a global or local minimum for gradient descent optimization [91].

3.12.5. Ridge Classifier

One of the most basic regularization strategies is to analyze multiple regression multicollinearity data. One projected value from numerous regression models is combined with others in this model to achieve a specified level of accuracy [92]. RGD is a technique for analyzing multicollinear datasets (a high-dimensional matrix). This classifier adjusts the weights to lower them to a very tiny number.

3.13. TECHNIQUES FOR DEEP LEARNING ARABIC SENTIMENT ANALYSIS

3.13.1. Convolutional Neural Network

One of the latest algorithms of neural networks, a convolutional layer, a pooling layer, and a fully connected layer are the three layers of a CNN. In a convolutional layer, the filters will be applied to the input image or text (matrix) to find the features; then, in a pooling layer, the size of the image or the text matrix will be decreased, but in the case of saving the important features [9]. The outputs of the pooling layer will be in a flattening case to convert all matrixes into one-dimensional matrixes to facilitate entry to the neural networks that will connect all layers in a fully connected layer.

CNN has proven its efficiency in image processing; surprisingly, it is now used in NLP and text processing. The algorithm can deal with texts as it deals with images. When

it treats the image as being dots or pixels, it will treat the text as a set of vectors representing an input matrix with I x B, where I is the number of non-repeated words in the text and B is the embedding value. The filter's vertical convolution (moving) will be the overall rows of the input matrix. It is worth mentioning that size of the filter is NxB, where N represents the number of words that relate to each other, or we can call it N-gram, and B is equal to the columns count in the input matrix. After applying the filter that moved vertically on the input matrix, the result will be the features with Mx1 size where M= (I-N+1). At last, max pooling and softmax (multi-classification) or sigmoid (binary classification) will be applied.

3.13.2. Bidirectional Long Short -Term Memory

It is an idea developed from RNN, based on inserting the output into the input again. It is used with sequential data that depends on time, such as stock prices, weather, and speech, until it is said needs time, so the texts are considered successive data [93]. Nevertheless, the problem with RNN is the inability to remember distant words. So Hocheriter & Schmindhuber launched the LSTM in 1997 to solve this problem. Since the LSTM has a memory cell, a word can be stored there and retrieved when necessary. BiLSTM is a two-LSTM sequence processing model, with one taking input in a forward direction and the other in a backward way.

3.13.3. Feed Forward Neural Network

Feed-Forward Neural Network (FFNN) is the simplest artificial neural network. Signals can only move in one direction with FFNN, from input to output [94]. There are no feedback loops where the output of one layer does not affect another layer of that system. Simple networks that link inputs and outputs are frequently FFNN.

PART 4

METHODOLOGY

4.1. COLLECTING DATA

The Twitter platform is one of the most important SM platforms that people use to express their opinions [12], and it contains different trends worldwide every day. Regarding tourism, Arabs love Turkey because of its historical landmarks, beautiful nature, and moderate climate suitable for tourism. Therefore, the research focused on collecting tweets in Arabic related to the opinions of Arabs from different countries and different dialects in the field of tourism, politics, living, asylum, and art in Turkey.

Google Colaboratory, a cloud service from Google Research, was used. It is an IDE that allows users to write the source code in their editor and run it from the browser [95]. Specifically, it supports the Python programming language and is geared toward ML tasks, data analysis, educational projects, etc. Based on Jupiter Notebook, this service is hosted completely free with a Gmail account, requires no configuration, and does not have to download or install Jupiter.

The first step to collecting data was to create a Twitter API account, which enables developers to access tweets[96] that are used on a specific topic that raises controversy. Using Python (Tweepy library) and Twitter API, which provided us with consumer_key, consumer_secret, access_token, and access_token_secret, tweets related to Turkey were collected using several hashtags. The hashtags were (العرب_في_تركيا)Arabs in Turkey, (الاتراك_) Turks, ((تركيا Turkey, (الدراسة_في_تركيا) Study in Turkey, and (المسلسلات_التركية) Turkish series. Also, the geographical region (Turkey) has been taken into account. However, we had difficulty with some Twitter users turning off geo-location sharing to preserve privacy. Also, we had difficulty dealing with the CSV file because the tweets were written in Arabic, which led to it being encrypted when reopening the file after closing it, as shown in Figure 4.1. So, we made

several copies of each file after getting the tweets, then combined all the files into one file using CMD instruction (copy mergerd.csv), and we made several copies of the file after merging to obtain 17000 tweets. Moreover, after checking and deleting duplicate tweets, advertisements, and tweets not relevant to the topic of Turkey, we reached a data size of 3136 tweets stored in a CSV file.

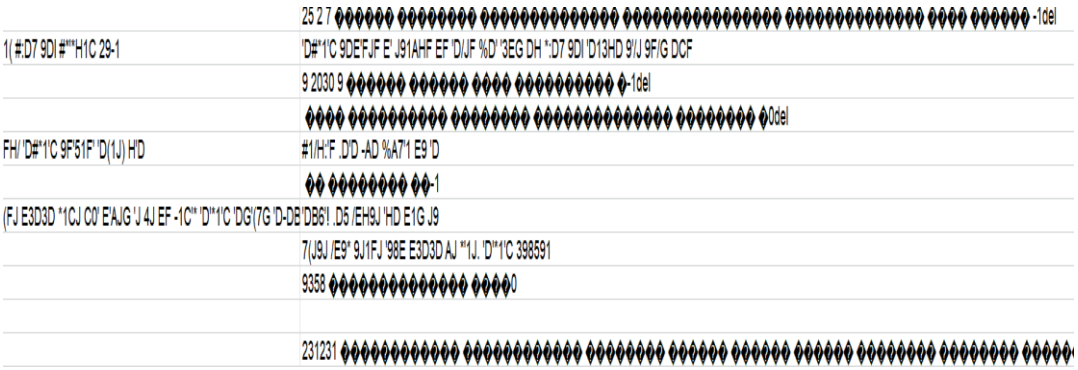

Figure 4.1. Encrypted the Arabic data.

Table 4.1. gives descriptive details regarding the datasets.

Table 4.1. Description of statistical information on the datasets used.

The total count of collected data	17000
The count of labeled tweets after PP	1336
Data set type	Arabic multi-dialect data (MSA, Egyptian dialect, Iraqi dialect, Saudi dialect, and Levantine dialect)
Data set domain	Multi-domain (economy, Turkish series, tourism, politic)
The count of words	198134
The maximum length of all tweets	51
The way of labeling	Hand-crafted
The count of positive tweets	1583
The count of negative tweets	1553
The number of unique words	11528
The number of stop words	8283

4.2. DATA LABELING

In order to achieve the task of analyzing textual feelings, it must be ensured that the sentence is subjective and contains feelings. In contrast, every objective sentence that contains the content of natural facts must be eliminated. So, after deleting the recurrent tweets. Ads, and non-relative tweets to Turkey, we started to read each sentence and label it as positive or negative by hand according to its context, after being sure that it is a subjective sentence. Thus, we got 1583 positive tweets and 1553 negative ones

Table 4.2. Splitting the dataset into training and testing.

The total dataset is 3136 tweets	
2508 tweets for training	628 tweets for testing

4.3. DATA PRE-PROCESSING

The main objective of the PP is to extract the most important information from the text [97]. These PP techniques assist convert noise from high-dimensional features to low-dimensional space.

About 17,000 tweets were collected using several hashtags. After that PP step started as follows:

Duplicated tweets were removed using Python code. However, duplicate tweets were not deleted due to a slight difference in the length of the duplicate tweet from the original tweet, so they have manually deleted. Also, Ads and tweets are not related to our work. In addition, the objective sentences have been deleted manually, and thus 3136 tweets were obtained about Turkey about living, tourism, art, and study. Figure 4.2 shows an example of our data before processing.

1	2022-03-31 15:4	RT @mstshar_trabzon: #رسالة_اليوم و #معلومه #ارتفين Artvin تقع شمال شرق #تركيا على ساحل البحر الأسود احتضنت هذه المنطقة شعوبًا وحضارات…
2	2022-03-31 10:4	@haya3laa تعالي لتركيا .. يا دبي تركيا جنة الله في الأرض منين ماتلتفتين حدايق وجبال وبحيرات وبحار وانهار
3	2022-03-30 21:2	RT @mstshar_trabzon: #رسالة_اليوم و #معلومه #ارتفين Artvin تقع شمال شرق #تركيا على ساحل البحر الأسود احتضنت هذه المنطقة شعوبًا وحضارات…
4	2022-03-30 14:0	الأطباق التركية تتصدر موائد نجوم السينما والمنتجين في حفل توزيع جوائز الأوسكار، بمبادرة من وزارة الثقافة والسياحة و… https://t.co/YZPuVwqgvR
5	2022-03-30 13:1	شركة تيمسا Temsa التركية تعزز مكانتها في السوق الفرنسية من خلال تسويق حافلات "لدينا أكثر من 5… https://t.co/ZPH7bqIFb8
6	2022-03-30 11:1	مسرح أسبيندوس وهو من أهم المسارح الرومانية القديمة تقام فيه العديد من المهرجانات #اسطنبول… https://t.co/JkJilWOVht

Figure 4.2. Screenshot of the data before processing.

For PP, operations include removing extra spaces between strings, editing numerical expressions, and checking for spelling errors, as well as removing misused punctuation marks [99], hashtags, years, URLs, usernames, RT (retweet) & cc (carbon copy). Also, Arabic writing techniques such as tashkeel (diacritics), repeated letters that are used to emphasize emotion as (كتيييييير حلو<< كتير حلو), and Tatweel (جميلـــــــــة << جميلة) was removed. This task helps to ensure consistency between

Moreover, the next stage of PP was as follows:

Stop words are the common words that are widely used and must be removed from the data. Considering that their existence or absence has no bearing on the sentence's meaning [9] and that they are words without emotional connotations, removing them will not change the SA but will hasten to process and lower the score [83].

Tokenization: It is the process in which a sentence is divided into many words called tokens [74], [83]. These tokens support the NLP model's construction or the context's understanding[84]. Tokenization assists in deciphering the text's meaning by analyzing

the word order. Moreover, for getting tokens, the TextBlob library [98] with words function was used, as shown in the example in figure 4.3:

```
from textblob import TextBlob

text=" وانا احب تركيا واحب الشعب التركي"

zen = TextBlob(text)

print(type(zen))
tokens = zen.words
print(tokens)
```

```
<class 'textblob.blob.TextBlob'>
['وانا', 'احب', 'تركيا', 'واحب', 'الشعب', 'التركي']
```

Figure 4.3. Screenshot of tokenizing text.

Normalization: in this step, some letters were replaced with another letter [99].

ى was converted to ي

إ أ آ ا was converted to ا

ؤ ئ was converted to ء

ة was converted to ه

4.4. FEATURE EXTRACTION

We used two feature extraction methods; the first was the Word2Vec method. Using the Word2Vec method from the gensim library, we converted each word to a vector with a length equal to (1, 300) and then used this vector as input to our DL models. Consequently, we also used the Arabert method to convert each word into a fixed-length vector equal to (1, 768). Exactly, we used the AraBERTv0.2-Twitter-base version [100].

4.5. SENTIMENT ANALYSIS MODELS

We used DL models to classify the data positively or negatively. Figure 4.4 shows suggested approach:

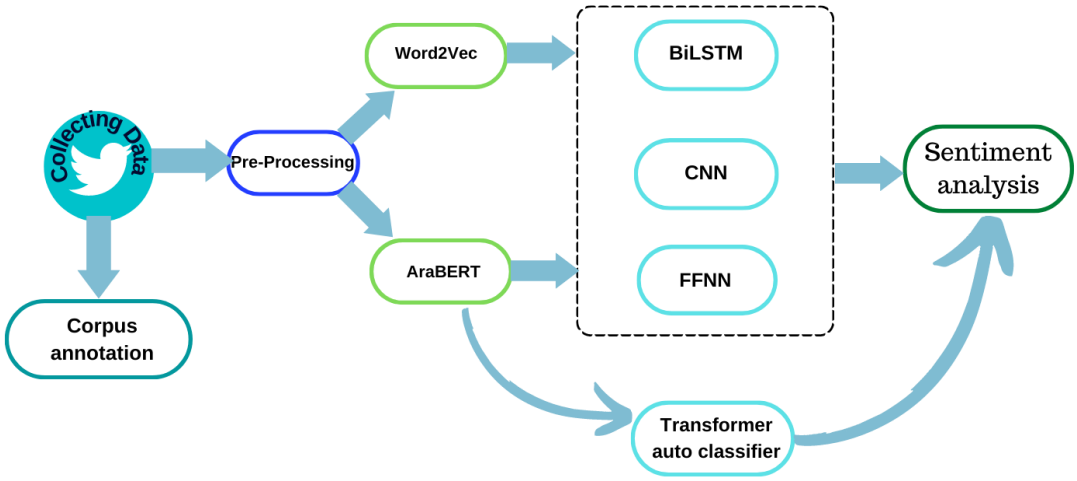

Figure 4.4. Pipeline of suggested ASA approach.

4.5.1. Bidirectional Long Short-Term Memory Model

Two LSTMs make up the sequence processing model; input is sent to one in a forward direction and the other backward.

The first step is PP, cleaning the data and extracting features; then, we build the model. So, after cleaning the data, we should split it into training & testing sets which the percentage of splitting sets is 20% for the testing set and 80% for the training set with a random state =42. Additionally, 50 samples from the stratified 5-fold cross-validation were employed.

A dense layer as output was applied, and a sigmoid function was suitable when we had binary classification. We assigned an adaptive moment estimation "Adam" optimizer. Adam is an optimization technique explicitly created for deep neural network training [101]. And a loss of type "binary_crossentropy."

Table 4.3. Hyperparameters of the BiLSTM model.

BiLSTM model training hyperparameters	**Word Representation Models**	
	Word2Vec	**AraBERT**
Word vector size	(1, 300)	(1, 768)
BiLSTM output dimension	64	512
Dropout rate	0.5	
Batch size	64	32
Optimizer	Adam	
Loss function	binary crossentropy	
Output activation function	Sigmoid	

4.5.2. Convolutional Neural Network

A dense layer with one neuron and a sigmoid function was added as an output layer to classify tweets into positive or negative.

Table 4.4 shows the values of the used parameters in the CNN model

Table 4.4. Hyperparameters of the CNN model.

CNN model training hyperparameters	**Word Representation Models**	
	Word2Vec	**AraBERT**
Word vector size	(1, 300)	(1, 768)
The number of output filter for convolution	32	
Kernel size for convolution	3	
Filter form	1 d	
Pooling method	Max pooling	
The count of neuron in fully connecting	250	512
Fully connected activation function	Relu	
Batch size	32	
Optimizer	Adam	
Loss function	binary crossentropy	
Output activation function	Sigmoid	

4.5.3. Feed Forward Neural Network

FFNN is a very simple type of artificial neural network. The information flows in one direction only.

Table 4.5. Hyperparameters of the FFNN model.

CNN model training hyperparameters	Word Representation Models	
	Word2Vec	AraBERT
Word vector size	(1, 300)	(1, 768)
Fully connecting neuron	16	64
Fully connected activation function	Relu	
Dropout rate	0.5	
Batch size	16	64
Optimizer	Adam	
Loss function	binary crossentropy	
Output activation function	Sigmoid	

4.5.4. Arabic Bidirectional Encoder Representations from Transformers with Auto Model for Sequence Classification

First, we must prepare our dataset. Then, we will load the AraBERT pretrained tokenizer and call it with our dataset. After that, build torch datasets and encode them. Then we move to the other last steps, using the Trainer class with our hyperparameters and Auto Model For Sequence Classification (Transformer auto classifier) from the transformers library.

AraBERT's Architecture:

Parameters: represent the number of learnable variables.

Transformer: represent the number of transformer blocks; the main aim is to transform a sentence (group of words) into a sequence of contextualized words (numerical representation).

Attention heads: these represent the size of the transformer block.

Hidden size is located between input and output, representing the mathematical functions, and assigning weights

Table 4.6. Hyperparameters used in AraBERT model with Transformer auto classifier.

Parameter	Value
optimizer	adam_epsilon =1e-8
learning_rate	2e-5
batch_size	32
warmup_steps	10
weight_decay	0.01

PART 5

RESULTS AND DISCUSSION

5.1. RESULTS

In this work, we presented a new collected and hand-crafted Arabic multi-dialect and multi-domain data set. We have read each sentence carefully and ensured that it includes Arab opinions and feelings about Turkey, then classify it as positive or negative according to context.

We use evaluation metrics to evaluate the textual data.

Four measures, precision (P), recall (R), f1 score (F), and accuracy (ACC), were employed in binary score in experimental investigations of the DL algorithm applied to ASA to assess how well they performed. A binary score is suitable for our case because we want to classify it into positive or negative. The following description provides information on the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) parameters used in these measures.

TP: Abbreviation of True Positive
TN: Abbreviation of True negative
FP: Abbreviation of False positive
FN: Abbreviation of False negative

P is used to compute the percentage of positive patterns that are correctly predicted out of all positive patterns:

$$\text{binary} - \text{P} = \frac{TP}{TP + FP} \tag{5.1}$$

R is used to compute the percentage of positives that are classified correctly:

$$\text{binary} - \text{R} = \frac{TP}{TP + FN} \tag{5.2}$$

In addition, F is given by:

$$\text{binary} - \text{F} = 2 \times \frac{\text{Precision} \times \text{Recal}}{\text{Precision} + \text{Recal}} \tag{5.3}$$

ACC computes the prediction's ratio correctly classified to the total instances evaluated.

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

Table 5.1 displays the results of the metrics (ACC, P, R, and F) for the DL models.

Table 5.1. Performance metrics of the suggested DL models.

WE Model	**Classifier**	**ACC**	**Binary-P**	**Binary-R**	**Binary-F**
Word2Vec	BiLSTM	0.77	0.75	0.83	0.79
AraBERT		**0.80**	0.78	0.81	0.80
Word2Vec	CNN	0.77	0.77	0.81	0.78
AraBERT		0.79	0.82	0.74	0.78
Word2Vec	FFNN	0.77	0.74	0.86	0.80
AraBERT		0.75	0.75	0.72	0.74
AraBERT	Transformer auto classifier	**0.90**	**0.91**	**0.89**	**0.90**

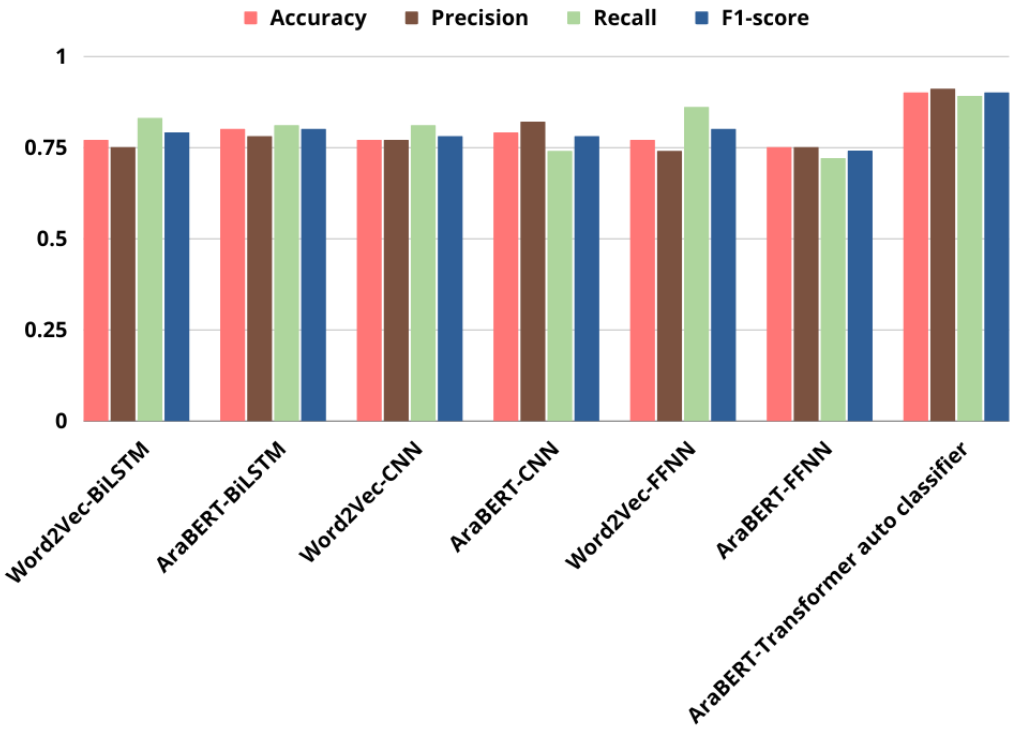

Figure 5.1. Bar chart of the metrics results of proposed DL algorithms.

5.2. DISCUSSION

This thesis's primary contribution was building a new Arabic corpus in various dialects and multi-domain. Table 5.2 shows examples of existing dialects in our dataset. Our work is vital because most previous studies focused on MSA or one dialect. The issue of collecting tweets containing feelings and opinions about Turkey, Turks, livelihood, and tourism was not an easy matter. The presence of many ads made the size of the data collected at the beginning large, but it was not valuable, so we had to manually delete the tweets that contained ads or did not contain feelings, which required a lot of time and effort. Then we had to convert the tweets into an annotated corpus and manually categorize them into positive or negative to ensure classification accuracy.

Table 5.2. Examples of dialects existing in the dataset.

MSA	توفر تركيا وسائل راحة خاصة للعرب الذين يعيشون ، لأن تركيا حرصت دائمًا على تزويد الأجانب بتعديلات جديدة لحماية حقوقهم في الجنسية ،
Egyptian dialect	الاتراك لو البطل والبطلة عاشوا في اخر الفيلم عادي ميعرفوش يعيشوا بجد لازم نتنكد وحد يموت بلا هدف
Iraqi dialect	مو قاسي علي كلش 😂 ما احب شاورما الأتراك مايحطون فيها لا طحينة ولا سلطة غثيثة
Saudi dialect	معيشة الأتراك مرة حلوة معاشرتهم سلسلة في الريف التركي
Levantine dialect	خالتي راحت تركيا وانبسطت كتير وانا حابة زور تركيا قريبا

The first step was cleaning the data and removing non-important words, symbols, URLs, or numbers. Then we discovered the data before and after the cleaning. We found that many non-important items, such as stop words and emojis, were deleted, saving time and facilitating the training process. Also, we tried stemming, and the Arabic language depends on suffixes and affixes. Therefore, language PP requires removing extra letters from words and returning the word to its root using the ArabicLightStemmer() class and light_stem function from the tashaphyne library. However, the result was not good enough because no suitable library can deal with the Arabic dialects. Figure 5.2 shows an example of stemming.

```
from snowballstemmer import stemmer
def stemming(text):
  ar_stemmer = stemmer("arabic")
  cleaned = list()#هي ليست عم نعبي فيها
  for word in text.split(" "):#عم امرق كلمة كلمة بعد ما اعمل سبليت بناء على الفراغ بين الكلمات
    stem = ar_stemmer.stemWord(word)#عم اعخد الجذر
    cleaned.append(stem)#ضيفو لليست
    #print(stem)
  return " ".join(cleaned)
stemming("أنا أحب العيش في تركيا لأن الاتراك نشيطون ويحبون السائحين وصميميون")
```

```
'انا احب عيش في تركي لان اتراك نشيط يحب سايح صميمي'
```

Figure 5.2 Screenshot of stemming text.

We used two different feature extractor WE methods, Word2Vec and AraBERT. To avoid overfitting, we added dense layers to the DL models with a relu activation function, especially since the data set is not so big and the model will face difficulty generalizing.

The architecture of the transformer makes it conceivable to parallelize training incredibly productively. Massive parallelization consequently makes it practical to prepare BERT on much information in a brief short period. Transformers utilize attention to notice connections between words It utilizes leveraging attention, a strong, profound learning calculation first found in PC vision models. Our proposition AraBERT to deal with ASA was a good way, primarily when we used the multi-dialect and multi-domain Arabic dataset. This was a new data set we collected from Twitter. A Transformer DL architecture with an attention mechanism that learns the link between words based on context underpins the success of AraBERT. WE models, such as Word2Vec, FastText, and GloVe, only consider one direction when generating WE representations for each word, whereas AraBERT transforms them into bi-directional contextual text representations that take into account both left and right contexts.

The critical finding is that AraBERT and Word2Vec WE, which take context into account in the inputs of BiLSTM, CNN, and FFNN models, contribute to ASA performance. Another finding is; When the ASA performances of the proposed approaches are analyzed, the text representations produced with the AraBERT WE have a higher capacity-rich representation capability than those produced with the Word2Vec WE. It is worth noting that using AraBERT as a representation of words gave higher efficiency than using Word 2Vec as input for the BiLSTM and CNN Models. Also, the best result that we obtained was using the automatic classifier that was built inside the transformer. This concentration exhibited that ASA has become one of the examination regions that have drawn the consideration of numerous analysts.

.

PART 6

SUMMARY

6.1. CONCLUSION

SA is a vital science that attracts researchers because of the diversity of aspects of life that need this science. SA is not limited to a specific language; it applies to many languages. However, the efficiency depends on the support of NLP for that language; for example, the libraries that support the English language are many and highly efficient, so the research on SA in the English language varied. On the other hand, treating the Arabic language is still a research gap and needs to be developed to reach high work efficiency.

This thesis presented a new ASA. The work began with developing a new dataset with a size of 3136 collected from Twitter, which contained Arabic tweets in the many colloquial dialects and MSA about Arab opinions about Turkey in the field of tourism, Turkish series, politics, and living. The tweets were then converted into an annotated corpus hand-crafted.

For extracting features, each word was represented as a vector according to two WE methods: Word2Vec and AraBERT. Then, three DL models were used for classification: BiLSTM, CNN, and FFNN. For each model, we used the Word2Vec vector and once AraBERT vector. In addition, we used the transformer auto classifier with AraBERT. The experiment showed that the transformer auto classifier achieved the highest performance compared with the DL classifiers. Also, AraBERT-BiLSTM outperformed the rest models.

6.2. FUTURE WORK

In future work, we want to focus on increasing the data set size and treating more dialects, then do SA according to many aspects. Also, we will apply ML algorithms and compare the results with our work. In addition, we want to apply the proposed model to other Arabic datasets.

REFERENCES

1. Karaoğlan, K. M., & Fındık, O. Extended rule-based opinion target extraction with a novel text pre-processing method and ensemble learning. ***Applied Soft Computing***, 118, 108524, (2022).

2. Hollander, J. B., Graves, E., Renski, H., Foster-Karim, C., Wiley, A., & Das, D. A (short) history of social media sentiment analysis. In Urban Social Listening ***Palgrave Macmillan***, London, (pp. 15-25) (2016).

3. Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. The history of social media and its impact on business. ***Journal of Applied Management and entrepreneurship***, 16(3), 79, (2011).

4. Peeters, M. M., van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. Hybrid collective intelligence in a human–AI society. ***AI & society***, 36(1), 217-238, (2021).

5. Ngiam, K. Y., & Khor, W. Big data and machine learning algorithms for health-care delivery. ***The Lancet Oncology***, 20(5), e262-e273. (2019).

6. Taboada, M. Sentiment analysis: An overview from linguistics. (2016).

7. Banić, L., Mihanović, A., & Brakus, M. Using big data and sentiment analysis in product evaluation. In 2013 36th ***International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)***, ***IEEE***. pp. 1149-1154, (2013).

8. Budiharto, W., & Meiliana, M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. ***Journal of Big data***, (2018).

9. Omara, E., Mosa, M., & Ismail, N. Deep convolutional network for Arabic sentiment analysis. In 2018 ***International Japan-Africa Conference on Electronics, Communications and Computations (JAC-ECC), IEEE***. pp. 155-159 (2018).

10. Ombabi, A. H., Ouarda, W., & Alimi, A. M. Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. ***Social Network Analysis and Mining***, 10(1), 1-13, (2020).

11. Bolbol, N. K., & Maghari, A. Y. Sentiment analysis of Arabic tweets using supervised machine learning. In 2020 ***International Conference on Promising Electronic Technologies (ICPET) IEEE***, pp. 89-93 (2020).

12. Al-Sorori, W., Mohsen, A. M., Ali, Y., Maqtary, N. A., Altabeeb, A. M., Al-Fuhaidi, B., ... & Al-Kaf, H. A. G. Arabic sentiment analysis towards feelings among COVID-19 outbreak using single and ensemble classifiers. ***In 2021 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE). IEEE***. pp. 1-6 (2021).

13. Alshammari, N. F., & AlMansour, A. A. Aspect-based sentiment analysis for Arabic content in social media. In 2020 ***International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*** pp. 1-6, (2020).

14. Elfaik, H. Deep attentional Bidirectional LSTM for Arabic sentiment analysis in Twitter. In 2021 1st ***International Conference on Emerging Smart Technologies and Applications (eSmarTA*) *IEEE***. pp. 1-8 (2021).

15. Alharbi, L. M., & Qamar, A. M. Arabic sentiment analysis of eateries' reviews: Qassim region case study. In 2021 ***National Computing Colleges Conference (NCCC) IEEE***. pp. 1-6, (2021).

16. Sayed, A. A., Elgeldawi, E., Zaki, A. M., & Galal, A. R. Sentiment analysis for Arabic reviews using machine learning classification algorithms. In 2020 ***International Conference on Innovative Trends in Communication and Computer Engineering (ITCE)***. ***IEEE***. pp. 56-63, (2020).

17. Abuuznien, S., Abdelmohsin, Z., Abdu, E., & Amin, I. Sentiment analysis for Sudanese arabic dialect using comparative supervised learning approach. In ***International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)***. ***IEEE***. pp. 1-6 (2020).

18. Alassaf, M., & Qamar, A. M. Aspect-based sentiment analysis of Arabic tweets in the education sector using a hybrid feature selection method. In 2020 14th ***International Conference on Innovations in Information Technology (IIT)***. ***IEEE***. pp. 178-185 (2020).

19. Das, A., Gunturi, K. S., Chandrasekhar, A., Padhi, A., & Liu, Q. Automated pipeline for sentiment analysis of political tweets. In 2021 ***International Conference on Data Mining Workshops (ICDMW) IEEE***. pp. 128-135 (2021).

20. Abdelli, A., Guerrouf, F., Tibermacine, O., & Abdelli, B. Sentiment analysis of Arabic Algerian dialect using a supervised method. In 2019 International ***Conference on Intelligent Systems and Advanced Computing Sciences (ISACS*) *. IEEE.*** pp. 1-6 (2019).

21. Dale, R. Natural language generation: The commercial state of the art in 2020. ***Natural Language Engineering***, 26(4), 481-487 (2020).

22. Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., & Kitchen, G. B. Natural language processing in medicine: a review. ***Trends in Anaesthesia and Critical Care***, 38, 4-9. (2021).

23. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. The Kaldi speech recognition toolkit. In ***IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF). IEEE Signal Processing Society***. (2011).

24. Aggarwal, C. C., & Zhai, C. A survey of text classification algorithms. In Mining text data (pp. 163-222). ***Springer***, Boston, MA. (2012).

25. Vijayarani, S., Ilamathi, M. J., & Nithya, M. Preprocessing techniques for text mining-an overview. ***International Journal of Computer Science & Communication Networks***, 5(1), 7-16. (2015).

26. Prabha, M. I., & Srikanth, G. U. Survey of sentiment analysis using deep learning techniques. In 2019 1st ***International Conference on Innovations in Information and Communication Technology (ICIICT)***. ***IEEE***. pp. 1-9. (2019).

27. JUGRAN, S., KUMAR, A., TYAGI, B. S., & ANAND, V. Extractive automatic text summarization using SpaCy in Python & NLP. In 2021 ***International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)***. ***IEEE***. (2021).

28. Deshmukh, R. D., & Kiwelekar, A. Deep learning techniques for part of speech tagging by natural language processing. In 2020 2nd ***International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) IEEE***. (2020).

29. Behdenna, S., Barigou, F., & Belalem, G.Document level sentiment analysis: a survey. EAI Endorsed Transactions on Context-aware Systems and Applications, 4(13), e2-e2. (2018).

30. Bhamare, B. R., Jeyanthi, P., & Subhashini, R. Aspect level sentiment analysis approaches. In 2019 5th ***International Conference On Computing, Communication, Control And Automation (ICCUBEA) IEEE***. (2019).

31. Alnawas, Anwar, and Nursal Arici. "The corpus based approach to sentiment analysis in modern standard Arabic and Arabic dialects: A literature review." ***Politeknik Dergisi*** 21.2 461-470 (2018).

32. Wise, E. K., & Shorter, J. D.. SOCIAL networking and the exchange of information. ***Issues in Information Systems***, 15(2), (2014).

33. Al-Otaibi, S., Alnassar, A., Alshahrani, A., Al-Mubarak, A., Albugami, S., Almutiri, N., & Albugami, A. Customer satisfaction measurement using sentiment analysis. International ***Journal of Advanced Computer Science and Applications***, 9(2) (2018).

34. Ramesh, B., & Weber, C. M. State-of-art methods used in sentiment analysis: A literature review. In 2022 ***Portland International Conference on Management of Engineering and Technology (PICMET) IEEE.*** (2022).

35. Almuqren, L., & Cristea, A. I. Predicting STC Customers' satisfaction using Twitter. ***IEEE Transactions on Computational Social Systems*** (2022).

36. Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. Evaluation of sentiment analysis in finance: from lexicons to transformers. ***IEEE*** access, 8, 131662-131682 (2020).

37. Saravanan, R., & Sujatha, P. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In 2018 Second ***International Conference on Intelligent Computing and Control Systems (ICICCS)***. ***IEEE***. (2018).

38. Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. ***Big data & society***, 3(1), 2053951715622512. (2016).

39. Nadeau, D., & Turney, P. D. A supervised learning approach to acronym identification. In ***Conference of the Canadian Society for Computational Studies of Intelligence. Springer***, Berlin, Heidelberg. (2005).

40. Xie, S., & Liu, Y. Improving supervised learning for meeting summarization using sampling and regression. ***Computer Speech & Language***, 24(3), 495-514. (2010).

41. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: ***springer***. (2009).

42. Sun, B., Zhang, Y., Zhou, Q., & Zhang, X. Effectiveness of semi-supervised learning and multi-source data in detailed urban landuse mapping with a few labeled samples. Remote ***Sensing***, 14(3), 648. (2022).

43. Zhou, Z. H. A brief introduction to weakly supervised learning. ***National science review***, 5(1), 44-53. (2018).

44. Wiering, M. A., & Van Otterlo, M. Reinforcement learning. Adaptation, learning, and optimization, ***springer*** 12(3), 729. (2012).

45. Hassan, A., & Mahmood, A. Deep learning approach for sentiment analysis of short texts. In 2017 3rd ***international conference on control, automation and robotics (ICCAR)***. ***IEEE***. (2017).

46. Alorini, D., & Rawat, D. B.. Automatic spam detection on gulf dialectical Arabic Tweets. In 2019 ***International Conference on Computing, Networking and Communications (ICNC) IEEE***. (2019).

47. Hamdi, A., Shaban, K., & Zainal, A. A review on challenging issues in Arabic sentiment analysis. (2016).

48. Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. Automatically assessing review helpfulness. In Proceedings of the 2006 ***Conference on empirical methods in natural language processing*** (pp. 423-430). (2006).

49. Gupta, S., Singh, R., & Singh, J. A hybrid approach for enhancing accuracy and detecting sarcasm in sentiment analysis. In 2020 IEEE ***International Conference on Computing, Power and Communication Technologies (GUCON)***. ***IEEE***. (2020, October).

50. Talafha, B., Za'Ter, M. E., Suleiman, S., Al-Ayyoub, M., & Al-Kabi, M. N. Sarcasm detection and quantification in Arabic tweets. In 2021 IEEE 33rd ***International Conference on Tools with Artificial Intelligence (ICTAI)***, ***IEEE***. (2021)

51. Wei, J., Liao, J., Yang, Z., Wang, S., & Zhao, Q. BiLSTM with multi-polarity orthogonal attention for implicit sentiment analysis. ***Neurocomputing***, 383, 165-173. (2020).

52. Izazi, Z. Z., & Tengku-Sepora, T. M. Slangs on social media: variations among Malay language users on Twitter. ***Pertanika Journal of Social Sciences & Humanities***, 28(1).

53. Aldayel, H. K., & Azmi, A. M. (2016). Arabic tweets sentiment analysis–a hybrid scheme. ***Journal of Information Science***, 42(6), 782-797. (2020).

54. Srivastava, R., & Bhatia, M. P. S. Challenges with sentiment analysis of on-line micro-texts. ***International Journal of Intelligent Systems and Applications***, 9(7), 31. (2017).

55. Srivastava, R., & Bhatia, M. P. S. Challenges with sentiment analysis of on-line micro-texts. ***International Journal of Intelligent Systems and Applications***, 9(7), 31. (2017).

56. Demiroz, G., Yanikoglu, B., Tapucu, D., & Saygin, Y. Learning domain-specific polarity lexicons. In 2012 IEEE 12th ***International Conference on Data Mining Workshops***, ***IEEE***. (2012).

57. Xing, F. Z., Pallucchini, F., & Cambria, E. Cognitive-inspired domain adaptation of sentiment lexicons***. Information Processing & Management***, 56(3), 554-564. (2019).

58. Sayed, A. A., Elgeldawi, E., Zaki, A. M., & Galal, A. R. Sentiment analysis for arabic reviews using machine learning classification algorithms. In 2020 International ***Conference on Innovative Trends in Communication and Computer Engineering (ITCE)***. ***IEEE***. pp. 56-63 (2020).

59. Shaalan, K., Siddiqui, S., Alkhatib, M., & Abdel Monem, A. Challenges in Arabic natural language processing. In Computational linguistics, speech and image processing for Arabic language (pp. 59-83). (2019).

60. D. Gamal, M. Alfonse, E.-S. M. El-Horbaty, and A.-B. M. Salem, “Modern education and computer science,” ***Modern Education and Computer Science***, vol. 1, pp. 33–38, (2019).

61. Zaidan, Omar F., and Chris Callison-Burch. "Arabic dialect identification." ***Computational Linguistics 40.1*** 171-202. (2014):

62. Duwairi, R., & Abushaqra, F. Syntactic-and morphology-based text augmentation framework for Arabic sentiment analysis. ***PeerJ Computer Science***, 7, e469. (2021).

63. Ariss, O. E., & Alnemer, L. M. Morphology based Arabic sentiment analysis of book reviews. In ***International Conference on Computational Linguistics and Intelligent Text Processing*** (pp. 115-128). ***Springer***, Cham. (2017).

64. Itani, M., Roast, C., & Al-Khayatt, S. Corpora for sentiment analysis of Arabic text in social media. In 2017 8th ***international conference on information and communication systems (ICICS). IEEE***. pp. 64-69 (2017).

65. Itani, M., Roast, C., & Al-Khayatt, S. Corpora for sentiment analysis of Arabic text in social media. In 2017 ***8th international conference on information and communication systems*** (ICICS) . ***IEEE***. pp. 64-69 (2017).

66. Duwairi, R. M. Sentiment analysis for dialectical Arabic. In 2015 ***6th international conference on information and communication systems (ICICS) IEEE***. (2015).

67. NASSR, Z., Nawal, S. A. E. L., & BENABBOU, F. Generate a list of stop words in moroccan dialect from social network data using word embedding. In 2021 ***International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)***. ***IEEE***. pp. 66-73 (2021).

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

69. Al-Saqqa, S., Awajan, A., & Ghoul, S. Stemming effects on sentiment analysis using large arabic multi-domain resources. In 2019 Sixth ***International Conference on Social Networks Analysis, Management and Security (SNAMS) IEEE***. (2019).

70. Tiwari, P., Mishra, B. K., Kumar, S., & Kumar, V. Implementation of n-gram methodology for rotten tomatoes review dataset sentiment analysis. In Cognitive analytics: Concepts, methodologies, tools, and applications (pp. 689-701). ***IGI Global***. (2020).

71. Huda, A. F., Ratnawulan, E., & Gumelar, D. R. Arabic part of speech (pos) tagging analysis using bee colony optimization (BCO) algorithm on Quran corpus. In 2021 7th ***International Conference on Wireless and Telematics (ICWT)***. ***IEEE***. (2021).

72. Zhang, Y., Jin, R., & Zhou, Z. H. Understanding bag-of-words model: a statistical framework. ***International journal of machine learning and cybernetics***, 1(1), 43-52. (2010).

73. Murali, S. R., Rangreji, S., Vinay, S., & Srinivasa, G. Automated NER, sentiment analysis and toxic comment classification for a goal-oriented chatbot. In 2020 Fourth ***International Conference on Intelligent Computing in Data Sciences (ICDS) IEEE***. pp. 1-7. (2020).

74. Sun, S., Luo, C., & Chen, J. A review of natural language processing techniques for opinion mining systems. ***Information fusion***, 36, 10-25. (2017).

75. Nagoudi, E. M. B., Ferrero, J., Schwab, D., & Cherroun, H. Word embedding-based approaches for measuring semantic similarity of Arabic-English sentences. In ***International Conference on Arabic Language Processing*** (pp. 19-33). ***Springer***, Cham. (2017).

76. Kurata, H., Tsukiyama, S., & Manavalan, B.. iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model. ***Briefings in Bioinformatics***, 23(4), bbac265. (2022)

77. Mohammed, Y., Srinivasan, S., Iyer, S., & Nagarajan, A. Defense of the Ancients (DOTA 2)-Draft Recommendation System. In 2022 6th ***International Conference on Trends in Electronics and Informatics (ICOEI)***. ***IEEE***. pp. 13-17 (2022).

78. Fadhil, I. M., & Sibaroni, Y. Topic classification in indonesian-language tweets using fast-text feature expansion with support vector machine (SVM). In 2022 ***International Conference on Data Science and Its Applications (ICoDSA) IEEE.*** (2022).

79. Santos, I., Nedjah, N., & de Macedo Mourelle, L. Sentiment analysis using convolutional neural network with fastText embeddings. In 2017 ***IEEE Latin American conference on computational intelligence (LA-CCI) IEEE.*** (2017).

80. Tan, K. L., Lee, C. P., Lim, K. M., & Anbananthen, K. S. M. Sentiment analysis with ensemble hybrid deep learning model. ***IEEE*** Access. (2022).

81. Pennington, J., Socher, R., & Manning, C. D. Glove: Global vectors for word representation. In Proceedings of the 2014 ***conference on empirical methods in natural language processing (EMNLP)*** pp. 1532-1543. (2014).

82. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is all you need", ***Advances In Neural Information Processing Systems***, 2017-Decem: 5999–6009 (2017).

83. KARAOĞLAN, K. M. "Özellik tabanli görüş madenciliğinde yapay zeka teknikleri kullanarak görüş hedefi çikarimi ve kategori tespiti" (Doctoral dissertation), (2022).

84. Shah, S. M. A., & Ou, Y. Y. TRP-BERT: Discrimination of transient receptor potential (TRP) channels using contextual representations from deep bidirectional transformer based on BERT. ***Computers in Biology and Medicine***, 137, 104821. (2021).

85. Matrane, Y., Benabbou, F., & Sael, N. Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case. In 2021 ***International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA) IEEE***. (2021).

86. Haddi, E., Liu, X., & Shi, Y. The role of text pre-processing in sentiment analysis. ***Procedia computer science***, 17, 26-32. (2013).

87. Misra, S., Li, H., & He, J. Noninvasive fracture characterization based on the classification of sonic wave travel times. Machine Learning for Subsurface Characterization, 243-287. (2020).

88. Jiang, L., Wang, S., Li, C., & Zhang, L. Structure extended multinomial naive Bayes. ***Information Sciences***, 329, 346-356. (2016).

89. Soni, K., & Yadav, P. Comparative analysis of Rotten Tomatoes movie reviews using sentiment analysis. In 2022 6th ***International Conference on Intelligent Computing and Control Systems (ICICCS)***. ***IEEE***. pp. 1494-1500 (2022).

90. Rafeek, R., & Remya, R. Detecting contextual word polarity using aspect based sentiment analysis and logistic regression. In 2017 IEEE ***International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) IEEE***. (2017).

91. Salinca, A. Business reviews classification using sentiment analysis. In 2015 17th ***International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) IEEE***. (2015).

92. Bashir, E., & Bouguessa, M. Data mining for cyberbullying and harassment detection in Arabic texts. ***International Journal of Information*** Technology and Computer Science, 13(5), 41-50. (2021).

93. Hochreiter, S., & Schmidhuber, J. Long short-term memory. ***Neural computation***, 9(8), 1735-1780 (1997).

94. Sundermeyer, M., Ney, H., & Schlüter, R. From feedforward to recurrent LSTM neural networks for language modeling. ***IEEE***/ACM Transactions on Audio, Speech, and Language Processing, 23(3), 517-529. (2015).

95. Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. Development of video-based emotion recognition using deep learning with Google Colab. ***TELKOMNIKA (Telecommunication Computing Electronics and Control)***, 18(5), 2463-2471. (2020).

96. Trupthi, M., Pabboju, S., & Narasimha, G. Sentiment analysis on twitter using streaming API. In 2017 IEEE 7th ***International Advance Computing Conference (IACC)***. ***IEEE***. pp. 915-919 (2017).

97. Anandarajan, M., Hill, C., & Nolan, T. The fundamentals of content analysis. In Practical text analytics (pp. 15-25). ***Springer***, Cham. (2019).

98. Kulkarni, A., & Shivananda, A. Exploring and processing text data. In Natural language processing recipes (pp. 31-62). ***Apress***, Berkeley, CA. (2021).

99. Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. Preprocessing Arabic text on social media. ***Heliyon***, 7(2), e06191. (2021).

100. Alami, H., Benlahbib, A., & Alami, A. High tech team at semeval-2022 task 6: intended sarcasm detection for Arabic texts. In Proceedings of the 16th ***International Workshop on Semantic Evaluation (SemEval-2022)*** pp. 840-843. (2022).

101. Qaisar, S. M. Sentiment analysis of IMDb movie reviews using long short-term memory. In 2020 2nd ***International Conference on Computer and Information Sciences (ICCIS) IEEE***. (2020).

RESUME

İnas CUMAOĞLU she graduated first, elementary, and high school education in this city then obtained a bachelor’s degree from Damascus university / Computer Engineering in 2014. She moved to İstanbul in 2015. Then in 2021 she started her master's education in the Department of Computer Engineering at Karabuk University