LARGE-SCALE ARABIC SENTIMENT CORPUS AND LEXICON BUILDING FOR CONCEPT-BASED SENTIMENT ANALYSIS SYSTEMS

KAVRAM-TABANLI DUYGU ANALİZİ SİSTEMLERİ İÇİN BÜYÜK ÖLÇEKLİ ARAPÇA DUYGU DERLEMİ VE SÖZLÜĞÜ OLUŞTURULMASI

Ahmed NASSER

Prof. Dr. Hayri SEVER

Supervisor

Submitted to Graduate School of Science and Engineering of Hacettepe University as a Partial Fulfillment of the Requirements for the Award of the Degree of Doctor of Philosophy in Computer Engineering

2018

This work named "**LARGE-SCALE ARABIC SENTIMENT CORPUS AND LEXICON BUILDING FOR CONCEPT-BASED SENTIMENT ANALYSIS SYSTEMS**" by **Ahmed NASSER** has been approved as a thesis for the degree of **DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING** by the below mentioned Examining Committee Members.

Prof. Dr. Erdoğan DOĞDU
Head ....

Prof. Dr. Hayri SEVER
Supervisor ....

Prof. Dr. Hasan OĞUL
Member ....

Assoc. Prof. Dr. Süleyman TOSUN
Member ....

Asst. Prof. Dr. Adnan ÖZSOY
Member ....

This thesis has been approved as a thesis for the degree of **DOCTOR OF PHILOSOPHY IN COMPUTER ENGINEERING** by Board of Directors of the Institute for Graduation School in Science and Engineering.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU
Directors of the Institute for Graduation School in Science and Engineering

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

☑ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**
(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, tezinin arama motorlarının önbelleklerinde kalmaya devam edebilecektir.)

☐ **Tezimin/Raporumun ………….. tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**
(Bu sürenin sonunda uzatma için başvuruda bulunmadığım taktirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı ve ya tamamının fotokopisi alınabilir)

☐ **Tezimin/Raporumun ………….. tarihine kadar erişime açılmasını istemiyorum, ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐ **Serbest Seçenek/Yazarın Seçimi**

04 / 01 /2018

Ahmed NASSER

ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate Studies in Science of Hacettepe University,

I declare that

- all the information and documents have been obtained in the base of the academic rules
- all audio-visual and written information and results have been presented according to the rules of scientific ethics
- in case of using others Works, related studies have been cited in accordance with the scientific standards
- all cited studies have been fully referenced
- I did not do any distortion in the data set
- and any part of this thesis has not been presented as another thesis study at this or any other university.

04/01/2018

Ahmed NASSER

ABSTRACT

LARGE-SCALE ARABIC SENTIMENT CORPUS AND LEXICON BUILDING FOR CONCEPT-BASED SENTIMENT ANALYSIS SYSTEMS

Ahmed NASSER

Doctor of Philosophy, Department of Computer Engineering

Supervisor: Prof. Dr. Hayri SEVER

January 2018, 120 pages

Within computer-based technologies, the usage of collected data and its size are continuously on a rise. This continuously growing big data processing and computational requirements introduce new challenges, especially for Natural Language Processing NLP applications. One of these challenges is maintaining massive information-rich linguistic resources which are fit with the requirements of the Big Data handling, processing, and analysis for NLP applications, such as large-scale text corpus. In this work, a large-scale sentiment corpus for Arabic language called GLASC is presented and built using online news articles and metadata shared by the big data resource GDELT. The GLASC corpus consists of a total number of 620,082 news article which are organized in categories (Positive, Negative and Neutral) and, each news article has a sentiment rating score value between -1 and 1. Several types of experiments were also carried out on the generated corpus, using a variety of machine learning algorithms to generate a document-level Arabic sentiment analysis system. For training the sentiment analysis models different datasets were generated from GLASC corpus using different feature extraction and feature weighting methods. A comparative study is performed, involving testing a wide range of classifiers and regression methods that commonly used for sentiment analysis task and in addition several types of ensemble learning methods were investigated to verify its effect on improving the classification performance of sentiment analysis by using different comprehensive empirical experiments. In this work, a concept-based sentiment analysis system for Arabic at sentence-level using machine learning approaches and a concept-based sentiment lexicon is also presented. An approach for generating an Arabic concept-based sentiment lexicon is proposed and done by translating the recently released English SenticNet_v4 into Arabic and resulted in producing Ar-SenticNet which contains a total of 48k of Arabic concepts. For extracting the concept from the Arabic sentence, a rule-based concept extraction algorithm called semantic parser is proposed and performed, which is generates the candidate concept list for an Arabic sentence. Different types of feature extraction and representation techniques were also presented and used for building the concept-based Sentence-level Arabic sentiment analysis system. For building the decision model of the concept-based Sentence-level Arabic sentiment analysis system a comprehensive and comparative experiments were carried out using variety of classification methods and classifier fusion models, together with different combinations of the proposed features sets. The obtained experiment results show that, for the proposed machine learning based Document-level Arabic sentiment analysis system, the best performance is achieved by the SVM-HMM classifier fusion model with a value of F-score of 92.35% and by the SVR regression model with RMSE of 0.183. On the other hand, for the proposed concept-based sentence-level Arabic sentiment analysis system, the best performance is achieved by the SVM-LR classifier fusion model with a value of F-score of 93.92% and by the SVM regression model with RMSE of 0.078.

Keywords: Arabic Sentiment Analysis; Concept-based Sentiment Analysis; Large-scale Corpus; Bigdata; Machine Learning; Ensemble Learning

ÖZET

KAVRAM-TABANLI DUYGU ANALİZİ SİSTEMLERİ İÇİN BÜYÜK ÖLÇEKLİ ARAPÇA DUYGU DERLEMİ VE SÖZLÜĞÜ OLUŞTURULMASI

Ahmed NASSER

Doktora Bilgisayar Mühendisliği

Tez Danışmanı: Prof.Dr. Hayri SEVER

Ocak 2018, 120 sayfa

Bilgisayar tabanlı teknolojilerinde toplanan verilerin kullanımı ve büyüklüğü sürekli artımaktadir. Bu sürekli artan büyük verinin işleme ve hesaplama gereksinimleri, özellikle Doğal Dil İşleme NLP uygulamalarında yeni bir zorluklar ortaya koymaktadır. Bu zorluklardan biri, Duygu Analizi (DA) gibi NLP uygulamalarında Büyük Verilerin ele alınma, işlenme ve analiz edilme gereksinimlerine uyan büyük ölçekli metin derlemi gibi zengin bir dilsel kaynağın sağlanmasıdır. Arapça dil için böyle büyük ölçekli bir kaynağın bulunmamasının zorluğu çözmek için, çevrimiçi haber Media'yı ve büyük veri kaynağı tarafından üretilen açık kaynak meta verilerini kullanarak inşa edilen GDELT büyük ölçekli Arapça duygu analiz derlemimizi (GLASC) tanıtmaktayız. GLASC derlimi, (Pozitif, Negatif ve Nötr) kategorilerinde düzenlenen toplam 620.082 haber makalesinden oluşmaktadır ve aynı zamanda, derlemimizdeki her haber makalesinin (-1 ve 1) aralığında bir duygu puanı vardır. Ayrıca, Makine öğrenme sınıflandırma ve regresyon yaklaşımlarına dayalı bir Arapça belge seviyesinde duygu analizi sistemi oluşturmak için GLASC derlemi kullanıp bazı deneyler gerçekleştirdik. Önerilen Makine öğrenmesi modellerini eğitmek için, farklı öznitelik çıkarma ve özellik ağırlıklandırma yöntemlerini kullanarak GLASC derlemimizden farklı veri kümeleri ürettik. Duygu analizi görevi için sıkça kullanılan sınıflandırma ve regresyon, yöntemlerinin testini içeren karşılaştırmalı geniş bir çalışma gerçekleştirilmiştir. Buna ek olarak, çeşitli kapsamlı deneyler kullanarak, duygu analizi için sınıflandırma performansının iyileştirilmesinin etkisini doğrulamak için, (Çuvallama, Yükseltme, Rasgele altuzay ve Öffekleme gibi) topluluk öğrenme yöntemlerinin çeşitli türleri araştırılmıştır. Bu çalışmada, makine öğrenme yaklaşımlarını ve kavrama dayalı bir duyugu sözlüğünü kullanarak, cümle düzeyinde Arapça için kavram tabanlı bir duygu analiz sistemi sunulmuştur. Yakın zamanda çıkan İngilizce SenticNet_v4'ü Arapça'ya çevirerek Arapça kavram temelli bir duygu sözlüğü üretmek için bir yaklaşım önerilmiştir. Üretilen Arapça konsept temelli duygu sözlüğü Ar-SenticNet toplam 48k Arapça kavram içermektedir. Arapça cümleden Konsepti çıkarmak için, anlamsal ayrıştırıcı olarak adlandırılan kural tabanlı bir kavramları çıkarma algoritması önerildi ve uygulanmıştır. Ayrıca, kavram tabanlı cümle düzeyinde Arapça duygu analizi sisteminin oluşturulması için farklı özellikler çıkarım ve gösterim teknikleri sunurak kullandık. Kavram tabanlı cümle düzeyinde Arapça duygu analiz sisteminin karar modeli oluşturmak için, farklı sınıflandırma yöntemi ve sınıflandırıcı füzyon modelleri kullanılarak, önerdiğimiz özellikler kümelerimizin farklı kombinasyonları ile kapsamlı ve karşılaştırmalı deneyler yapılmıştır. Elde edilen deney sonuçlarımıza dayanarak, önerilen Makine öğrenmesi tabanlı Doküman düzeyinde Arapça duygu analiz sistemimiz için, en iyi performans % 92.35 F-skoru değeri olan SVM-HMM sınıflandırıcı füzyon modeliyle ve 0.183 RMSE değeri olan SVR regresyon modeli ile, gerçekleştirilmiştir. Öte yandan, önerilen konsept tabanlı cümle düzeyinde Arapça duygu analiz sistemimiz için, en iyi performans, %93.92'lik bir F-skoru değerine sahip SVM-LR sınıflayıcı füzyon modeliyle ve 0.078 RMSE değeri olan SVR regresyon modeli ile, gerçekleştirilmiştir.

Anahtar Kelimeler: Arapça Duygu Analizi; Kavram Tabanlı Duygu Analizi; Büyük Ölçekli Derlem; Büyük Veri; Makine Öğrenmesi; Topluluk Öğrenimi

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor Prof. Dr. Hayri SEVER for his continuous support in my PhD. study, and also for his patience, motivation, enthusiasm, and immense knowledge.

Besides my supervisor I would like to express my thanks to my thesis committee members Prof. Dr. Erdoğan DOĞDU and Asst. Prof. Dr. Adnan ÖZSOY for their support, guidance and helpful feedbacks.

I would extend also my sincere gratitude to all who gave me much of their valuable time to help me during the research, especially Alaettin UÇAN.

Moreover, I would like to thank YTB and Iraqi MOHESR for sponsoring and funding my study.

My special thanks goes to my family for their patience, motivation and continuous support during my study.

Ahmed Nasser

CONTENTS

Page

FIGURES

TABLES

Page

ABBREVIATIONS AND SYMBOLS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ArSenL	Arabic sentiment analysis lexicon
Ar-SenticNet	Arabic SenticNet
Ar-WordNet	Arabic WordNet
AWATIF	A Multi-Genre Corpus for Modern Standard Arabic
BoW	Bag of Words Features
CAMEO	Conflict and Mediation Event Observations
CBF	Concept-based Features
CSV	Comma-Separated Values
DTIC	Defense Technical Information Center
D-Tree	Decision Tree
En-SenticNet	English SenticNet
En-SentiWordNet	English SentiWordNet
En-WordNet	English WordNet
En-WordNet	English WordNet
FCA	Formal Concept Analysis
FFCA	Fuzzy Formal Concept Analysis
GDELT	Global Database of Events, Language, and Tone
GKG	Global Knowledge Graph
GLASC	GDELT Large-Scale Arabic Sentiment Corpus
HAAD	Human Annotated Arabic Dataset
HMM	Hidden Markov Model
IMDB	Internet Movie Database
JSTOR	Journal Storage
KNN	K Nearest Neighbor
LABR	A Large-Scale Arabic Sentiment Analysis Benchmark

LB	Lexicon Based
LEX	Lexicon Based Features
LR	Logistic Regression
MAE	Mean Absolute Error
ME	Maximum Entropy
ML	Machine Learning
MLP	Multilayer Perceptron
MLR	Multilinear Regression
MSA	Modern standard Arabic
NB	Naïve Bayes
NLP	Natural Language Processing
OCA	Opinion Corpus for Arabic
OM	Opinion Mining
PoS	Part of Speech
PoST	Part of Speech Tag
RMSE	Root Mean Square Error
SA	Sentiment Analysis
SAMA	Standard Arabic Morphological Analyzer
SAMAR	Subjectivity and sentiment analysis for Arabic
SO	Semantic Orientation
SP	Semantic Parser
SQL	Structure Query Language
SSA	Subjectivity and Sentiment Analysis
SVM	Support Vector Machine
TF	Term Frequency
TF–IDF	Term Frequency-Inverse Document Frequency
W2V	Word2Vector Features

1. INTRODUCTION

1.1. Overview

The process of examining and identifying the sentiment or emotions that realis behind the words is called Sentiment analysis (SA). The main purpose of SA is to capture the tone of feeling that expressed by the words used within the text. The terms of Sentiment Analysis [1] and Idea Mining [2] first appeared in 2003. Elliott [3] and Ortony et.al. [4] carried out the primitive SA method which depends on effective words existence. Although SA consists of hybrid studies by means of combining the methods; it mainly consists of two methods: these methods are Machine Learning (ML) based methods [5] and Lexicon Based (LB) methods [6]. SA or often called opinion mining (OM) utilizes different methods for information extraction such as text analysis, natural language processing NLP, and computational linguistics [7]. SA or Opinion Mining (OM) is used in wide range of area such as; evaluation, social media marketing, and customer service. In general, SA aims to identify the attitude of the speaker/writer or sentiment polarity of textual contents for a particular title or subject.

There are many studies that deal with the automatic sentiment identification in the literature. The preliminary studies in SA include using dictionary-based ML methods [8].

Some of these studies have focused on using varies features together with different ML approaches for achieving the SA task. Kim and Hovy [9] proposed a method for extraction the word elements related to documents by using a sentiment dictionary. Dave et al. [2] introduced a method for capturing the syntactic properties of sentimental texts using bigram and trigrams. Agarwal et al. [10] used a dictionary contains predefined positive and negative words. Wilks and Stevenson [11] used a set of syntactic features or vocabulary types which also helps to eliminate the ambiguity.

There are also studies conducted by Aizawa [12], Scheffer and Wrobel [13], Serrano and Castillo [14] that focused on using different structures in order to represent the features which associated with a document, such as an event vector frequency.

On the other hand, the studies conducted by Joachims [15], Vapnik and Lerner [16] showed that using linear Support Vector Machine (SVM) with the obtained document attributes has achieved a very good performance in regard of text sentiment classification. In addition, Pang and Lee [5] investigated the use of graphical representations for SA in texts and proposed a concept for the of use n-grams with frequency vectors. ML classifiers such as Naive Bayes (NB), Decision Tree (D-Tree), K Nearest Neighbor (KNN) and Support Vector Machines (SVM) have been widely used for SA task [17].

The common approach for LB SA is done by using a dictionary consisting of words and sentiment polarities that associated with these words. Esuli and Sebastiani [18] have proved the effectiveness of using the SentiWordNet dictionary in the SA of text documents. SentiWordNet sentiment dictionary has been featured and used in many works such as; Product evaluation of Hamouda and Rohaim [19], news headlines of Chaumartin [20] and multilingual sentiment analysis studies of Denecke [21]. The most basic technique in applications performed by using a dictionary like SentiWordNet to collect the polarity scores of words in a document and then estimate the overall sentiment polarity based on the collected scores.

ML and LB approaches are also used in the SA researches related to the Arabic language in the literature. However, the number of these researches for Arabic is significantly small when it’s compared with the number of researches for other languages such as English.

Nowadays the massive and the rapid growth of the Big Data internet resources handling introduced a new set of difficulties especially in Artificial Intelligence applications such as NLP [22]. One of the important difficulties in such applications is maintaining large information-rich resources such as a large-scale text corpora which are considered as the most vital linguistic resources that can be used for training and evaluation many NLP ML applications such as SA [23].

In NLP applications large-scale resources become an essential demand for ensuring the performance and the robustness of these applications [24].

The importance of the corpus size with regard to the number of word in the corpus is investigated in [25], where the authors noticed that within a given corpus the appearance probability of a particular words follows the distribution that achieved with Zip's Law [26], which state that “Within a corpus the words occurrences frequencies tend to decrease in a quadratic-like manner.”

If we generated a list consist of all unique words within a certain corpus together with its corresponding occurrence frequencies, then sort this list descendingly based on the occurrence frequencies of the words. We can see that the last word in the list tends to appear two times lesser than the previous word in the list and so on. This can prove the relation between the corpus size and the number of words within the corpus. So that in the case of corpus size is being small, the probability of many words to be not appeared in this corpus is high, and vice versa.

There is a limited number of resources available that can be used for Arabic SA task in the literature. Table 1.1 provides a comparison between the popular Arabic data resources used in the most SA researches that available in the literature, in regard to the number of citations, the size of data, the source of data and the provided sentiment categories.

Table 1.1 A comparison between Arabic sentiment analysis data resources

Corpus / Dataset	Citations	Size	Data source	Categories
OCA [27]	118	500	Movie reviews	Positive Negative
Awatif [28]	80	2,855	Web forums, Wikipedia talk pages and Penn Arabic Treebank	Positive Negative Neutral
LABR [29]	39	63,000	GoodReads	1 to 5 rating
SAMAR(TGRD) [30]	139	3,015	Twitter	Positive Negative
SAMAR(THR) [30]	139	3,008	Wikipedia Talk Pages	Positive Negative
SAMAR(MONT) [30]	139	3,097	Arabic forums	Positive Negative
HAAD [31]	16	2,389	Book reviews	Positive Negative Conflict Neutral
Multi-domain Arabic Sentiment Analysis datasets [32]	26	32,338	Movies, hotels, restaurants and products reviews	Positive Negative

From Table 1.1, we can clearly see that these Arabic SA data resources are very limited in size. This lack of availability of the large-scale resource for the Arabic language has motived us to carry out this work by building a large-scale Arabic SA corpus using Online News Media and utilizing the metadata that provided by the bigdata resource GDELT [33].

Nowadays the social media and internet become a very simple and effective platform for the people for expressing their emotions and opinions through written text. The need of capturing the opinion of the public has raised due to the exceptional range of benefits that, include marking, business management, and financial forestation. However, mining opinion from languages is a very complex task because of it’s need to a deep and complete understanding of the rules of the language. Conventional SA approaches are mainly dependent on the parts of the text in which opinions are expressed, based on features such as words co-occurrence frequency, keywords, and terms polarity. However, because these syntactical approaches are not relying on the natural language semantic and effective information of the text, these approaches are not efficient in detecting complex emotions.

Concept-based approaches [35] are relying on the semantic and effective information that associated with the natural language opinions, which are represented as the concepts. Concept-based SA approaches utilize the semantic networks and web ontologies for analyzing the textual contents semantically.

This concept-based SA method is considered to be superior to other ordinary sentiment analysis methods because it's able to detect the emotions that conveyed by multi-word expressions concepts [34] [35]. Rather than gathering separated opinions, concepts based analysis enable a comparative fine grind feature-based analysis. Common and commonsense can be considered as the key that enables feature spotting and polarity detection and it also necessary for dismantling the language into sentiment. Approaches of concept based sentiment analysis emphasize the effective knowledge-based resources such as WordNet [36], SentiWordNet [18] and SenticNet [37] [38].

Since concept-based approaches for SA offered more advantages when they compared to traditional approaches, and since the concept-based approaches are not presented and used yet for Arabic SA according to our best knowledge, this motived and encouraged us for carrying out this work by presenting a concept-based SA system for Arabic using ML approaches.

1.2. Aims and Contributions

We can summarize the aims and contributions of this work as following.

1. Uses Arabic news metadata that provided by bigdata resource GDELT to generating the largest up-to-date resource for the Arabic language (GDELT Large-Scale Arabic Sentiment Corpus GLASC), which we believe it would help to improve not only SA

application but also a wide spectrum of NLP applications for the Arabic language in general.

2. Use our large-scale sentiment corpus to generate four datasets based on different feature extraction and feature weighting method. These datasets can be used for building and evaluate ML-based SA systems for the Arabic language.

3. Building a Document-level Arabic SA system based on ML classification and regression approaches, where a ML-based classifier model is used to assign an Arabic document into sentiment category in term of (positive, negative or neutral) and a ML-based regression model used for predicting the sentiment score of the Arabic document based on its sentiment orientation.

4. Carrying out various experiments on the datasets generated from our large-scale corpus, using ML algorithms to train a Document-level Arabic sentiment classifier. We have focused on using the ML classification and regression methods that widely used in SA works on the literature such as K-nearest neighbors (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), Neave Bayes (NB) and Hidden Markov Model (HMM) for building the sentiment classification model, and Support Vector Regression (SVR), Multilinear Regression (MLR) and Multilayer Perceptron (MLP) to build the sentiment score prediction model. Then conducting a comparative assessment of the performance of these different classification and regression methods base on using the different datasets generated from our large-scale corpus.

5. Verifying the ensemble learning effectiveness for sentiment classification task. We investigate the effectiveness of using popular classifier model ensemble techniques such as (Bagging, Boosting, and Random Subspace and staking) in enhancing the classification accuracy of the base learners such as (SVM, HMM, NB, NN, and KNN) for sentiment classification.

6. Generating a concept-based sentiment lexicon for Arabic (Ar-SenticNet) by translating the English version of the concept-based sentiment lexicon SenticNet

using two-way translation approach based on English-Arabic cross-language WordNet mapping and Google translation service.

7. Using Arabic WordNet to extend the translated Ar-SenticNet concept-based sentiment lexicon by adding extra senses to the concepts in Ar-SenticNet.
8. Building a concept-based SA system for Arabic sentence-level SA using our translated Arabic SenticNet concept based sentiment lexicon and ML approaches.
9. Presenting and utilizing of various feature extraction and representation techniques for building the concept-based sentence-level Arabic SA system. These techniques used to extract various feature sets from the input sentence, which used to build the ML decision model. These feature set are concept based features, lexicon based features, Bag of Word features and Word2Vector features.
10. Exploring the effectiveness of using several types of features combinations in improving the performance of the ML decision model that used for the concept-based Sentence-level Arabic SA system.

1.3. Thesis Outline

All the works that done within the scope of the thesis are explained in the following sections of the thesis. The structure of the thesis can be summarized as follows:

In Chapter 2, we explain and discuss the background information which allows the reader to understand the approaches that presented in the thesis. In Chapter 3, we presented in detail our proposed approaches that used for achieving the goals of this thesis. In Chapter 4, we presented different test experiments that applied for evaluating the performance of our proposed ML-based SA approaches and also provide a discussion and comparison of the obtained results. Finally, in Chapter 5 we presented the conclusion of this study followed by our proposed future works.

2. BACKGROUND

In this chapter, we review and discuss the fundamental information that helps to understand the approaches presented and used within this thesis scope. In Section 2.1, we described the approaches that commonly used for Arabic SA. brief definition of the Arabic language is provided in Section 2.2. In Section 2.3, we provide a literature review summary of the studies and researches that deal with Arabic language SA. In Section 2.4, we described the bigdata platform GDELT which is considered as a resource for building our large-scale SA corpus for Arabic. In Section 2.5 we reviewed and discussed the ML approaches which used within the scope of the thesis for our proposed SA systems. In Section 2.6, we described and discussed the concept-based approaches for SA and reviewed the data resources that used for building the concept-based SA and its applications.

2.1. Sentiment Analysis for Arabic Language

Currently, SA or opinion mining is considered as one of the most rapidly emerging research areas due to the immediate need of processing the opinionated web contents coming from social networks and web blogs. SA is the task of determining the sentiment polarity of textual contents i.e. SA determines whether the emotions that expressed by a specific piece of text, is positive, negative or neutral [39]. There are many supervised and unsupervised approaches in the literature that deals with the SA of the Arabic language which are used to achieve the SA task in Document-level or Sentence-level [40]. The supervised approach or the corpus-based approaches involves the generating of a sentiment decision model based on using an annotated sentiment corpus for training a different types of ML classification approaches such as K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (D-Tree), Support Vector Machine (SVM), and etc... The alternative unsupervised approach or LB approaches use a sentiment specific dictionaries in order to identify the polarity of a text based on the sentiment polarity of the individual words used in that text.

Hybrid approaches called semi-supervised are also available for SA, this hybrid approaches can be formed by combining both of ML and LB SA approaches [41]. It may be worth stating that Subjectivity and SA (SSA) has been receiving more attention among scholars [42] [30]. The SSA studies are similar to SA studies however, SSA based approaches are able to predict the subjective or objective classes of the text beside predicting the sentiment polarity [28].

2.2. Arabic Language

The Arabic language is considered as Semitic languages (the language that has a complex and uncommon morphology) and is mostly spoken in the North Africa and Middle East regions by an over 350 million people. The Arabic language is considered to be one among the five mostly spoken languages in the world and, one of the 10 most used languages on the internet [43]. The Arabic language has a 28 letters alphabet and its writing style is from right to left [44]. The Arabic Language has two forms: the first one is called Modern Standard Arabic (MSA) which is the formal language that commonly used in the media and literature in the Arab world. MSA is following the grammatical rules of Quran and consist of a vocabulary size greater than 1.5 million words. The second type called Dialectal Arabic or slang which is considered as the daily used language in Arab countries. Although Dialectal Arabic is driven from the MSA, it may feature some variations in vocabularies and grammatical rules depend on the dialect used in each country [45].

2.3. Literature Review

Although the Arabic language is considered as one of the mostly used languages on the internet, it has been taken less attention with regard to NLP researches especially SA, compared to other languages such as English [46]. This inadequately in researches of the SA for Arabic can be due to the complex structure and nature of the Arabic language and also the insufficiency of a quality linguistic resources that can be used for Arabic SA such as corpora and lexicons. Some of the important Arabic SA studies were summarized as following.

In [27] Rushdi-Saleh et al. proposed a document-level supervised SA approach. They generated an Arabic opinion corpus called OCA using the online movies reviews. For identifying the sentiment polarity, they used two types of ML classification methods which are NB and SVM. To extract the features from the Arabic documents they used various feature extraction methods based on n-gram representations and two different feature weighting techniques based on “Term Frequency” (TF) and “Term Frequency-Inverse Document Frequency” (TF-IDF).

Shoukry and Rafea [47] used sentence-level supervised SA approach for the Arabic language by collecting the required data for SA from Twitter. They applied two different feature extraction methods based on using Bigrams and Unigrams and TF weights together with NB and SVM ML-based classifier for building their proposed approach.

In [48] Mountassir et al. three different solutions were proposed for solving the imbancaing issue in the datasets that used for SSA. These methods include; “eliminate by clustering”, “eliminate similar”, and “eliminate farthest”. In addition to that, they built a supervised approach for Document-level Arabic SA based on different types of ML classification methods such as KNN, NB, and SVM. They used a binary weighting which is based on term presence where the documents are considered as bags-of-words. Two types of imbalanced of Arabic and English corpus were used for evaluating their system, the first one consists from Arabic movie reviews that collected from “Al-Jazeera’s website” and the second one consists from English product reviews and collected from the SINAI.

Ahmed et al. [49] presented several solutions for addressing the challenges in Arabic SSA subject. They used Sentence Level Supervised SA on data collected from Twitter. They investigated various types of ML classifiers such as D-tree, NB, and SVM which are used for identifying the sentiment polarity of an input tweet. They also investigated the effectiveness of using different preprocessing approaches that used for extracting and reducing the features.

In the work presented by Abdulla et al. [41] a Sentence-level SA system is presented to be used for Arabic Twitter. The proposed SA system involves using two different ML-based and LB-based SA approaches. They carried out different experiments on different types of ML classifiers such as KNN, NB, D-tree, and SVM for building an ML-based SA tool which aims to identify the sentiment polarity of the Arabic text. The lexicon is built and extended in three stages and at each stage, they measured how is the size of the lexicon is the impact on the accuracy of their proposed method. For the LB approach, they used their sentiment lexicon for extracting the sentiment terms with the corresponding sentiment score from the Arabic text and the final polarity of the text is found by summing out the sentiment scores of the extracted terms. They also performed a comparison between the two ML-based and LB-based SA approaches in regard to accuracy and performance.

Abdulla et al. [50] proposed Sentence-level supervised SA system for Arabic that built using a manually annotated large dataset. The dataset is collected from Arabic social network and used together with two different ML classification methods (NB and SVM) for building the supervised SA system. The authors also considered using different additional information related to the reviews/comments in their SA system, such as the number of likes and the gender of the writer. Finally, they conducted a comprehensive experiment in order to analyze the performance of their proposed approach.

In [51] Elmasry et al. presented a sentence-level supervised SA approach for Arabic. The dataset was collected from different Arabic news websites. The authors are also built a sentiment lexicon consist of the Arabic opinion idioms and slang words. Each entry in this lexicon contains two classes: satisfaction and dissatisfaction classes. They used their proposed SA approach on Facebook for classifying the comment that related to Arabic news by using SVM classification method based on the Gaussian kernel. They tested several types of methods for comment classification task in their proposed approaches these methods based on using either a lexicon consists of classical sentiment words or their idioms and slang words lexicon.

El-Makky et al. [52] presented a new sentiment lexicon for Arabic which built by combining two Modern Standard Arabic MSA lexica, namely, MPQA [54] and ArabSenti [55] with two Egyptian Arabic lexica built from Twitter. They used both the Sentence-level Supervised and Unsupervised SA. For the Semantic Orientation (SO), they proposed an augmented LB which depends on the presence of the sentiment words (looked-up from a sentiment lexicon). These words expressed positive or negative sentiments. The sentiment of the tweet that results from the modified algorithm was used as a (SO) score which was a component of the proposed feature vector. Feature vector consisted of: “Semantic Orientation feature”, “Tweet specific features”, “Language independent features”, “Stem level features”, and “Normalized word feature”. An ML-based SVM classifier is used as a subjectivity and polarity classifier.

Authors in [40] investigated both of LB and ML-based SA approaches for building a system for Arabic SA at the document and sentence levels. They used online Arabic movie reviews for generating their sentiment lexicon and dataset. They introduced a feature extraction approach based on the grammar structure of Arabic sentences to extract features such as (objects, adjectives, phrase type, verbs, and subjects) and used it together with the sentence sentiment polarity that obtained by LB approach, for generating the input feature vectors for SVM classifier. On the other hand, document-level SA approach is done by partitioning the input document into different chunk then calculating the positive, negative, and neutral sentence ratio at each chunk and use it as input to SVM classifier.

In [32] a Document-level weakly supervised Arabic SA was done. They collected three datasets from different domains such as education, politics, and sport for Arabic language and used them for sentiment lexicon generation. They used LB SA method for identifying the sentiment polarity labels of a set of Arabic documents and used them together as a dataset for training a Maximum Entropy (ME) classifier which in turn used for identifying the sentiment polarity labels of another set of Arabic documents that used for training KNN classifies.

Yet, a Document-level Unsupervised SA for Arabic is also used in [53]. They used a pattern recognition semi-supervised approaches with the “Conditional Random Fields” (CRF) feature analysis technique. The data collected from News articles from Arabic Language Technology Center "ALTEC". For obtaining Arabic strongly and weakly subjectivity clues, they manually translated the MPQA subjectivity lexicon into Arabic and marked the polarity and strength of each Word.

In order to achieve the best performance, they compared and combined three various models such as;

1. Opinion sources identification using traditional pattern matching by using key phrases, and POS tags.
2. Opinion sources identification using sequential tagging CRF classifier.
3. Opinion sources identification using sequential tagging CRF classifier with the use of patterns as a feature.

The features used in this work are: “The Semantic Field (SF) Feature”, The Word and Its Surrounding”, “Part of Speech Tag (PoST) Feature”, “The Named Entity Features”, “Base Phrase Chunk (BPC) Feature”, “Pattern Feature”, “Strong and Weak Subjectivity Clue Features”, “Subjectivity Classifier Feature” and “Objectivity Classifier Feature”.

In [28] a Sentence-level Supervised SSA approach for Arabic social media was considered. An SSA ML-based approach for Sentence-level Arabic SA is built and used for social media. They built a sentiment corpus by collecting different Arabic texts from various social media websites and labeling them manually. They used this corpus for building a subjectivity and sentiment classifier based on SVMlight classification algorithm. They also concentrated on the adjectives by considering them as a separated feature that associated with the words presences in the feature vector.

In [54], a domain-specific sentiment lexicon for Arabic is built and used for creating an LB Arabic Twitter SA system at Sentence-level. For obtaining the tweet polarity they used to approaches, the first one is done by aggregating the sentiment polarity weights of each term found in the tweet. The second approach is taken into the consideration both of the negative and positive weights for each term in the tweet and called double polarity (DP).

In [55] Al-Kabi et al. a SA tool is built for Arabic based on Sentence-level Unsupervised SA. The Data are collected from the Arabic social media reviews and comments. This dataset was used to create three polarity dictionaries: (Arabic, English, and Emoticons). These dictionaries were used to empirically evaluate SocialMention and Twendz. A program was designed and implemented to encode the contents of the three polarity dictionaries. This program starts reading the dictionary contents and assigns to each entry in this dictionary one of the following three values: (1-positive, 0-negative and ?-neutral). Each dictionary entry either uses Arabic, English, or Emoticons. After identifying the polarity of each entry in the polarity dictionaries, the program starts reading and determining the polarity of each entry (comment or review) in the collected datasets, by creating a sequence of symbols (0, 1, ?) to determine the final polarity of each entry in datasets.

Duwairi et al. [56] used a supervised SA approach for tweets in the Arabic language. The authors generate a large dataset form tweeter and Facebook comments in different domains and manually tagged the polarity for each tweet and comment in the dataset. they used three different ML-based classifiers such as NB, KNN, and SVM, as sentiment classification method.

Duwairi et al. [57] are also proposed a supervised learning approach for SA of tweets written using Arabizi (writing Arabic using Latin letters). They used rule-based method for converting arabizi tweets to Arabic. Then the using crowdsourcing for assigning the sentiment polarity to each tweet to generate the dataset which used to build SA framework using two different classification techniques such as NB and SVM.

These Arabic SA researches that previously described are summarized in Table 2.1 in form of the used data recourses, the size of the data used, the type of dialect, the approach used for SA, SA level and advantages and disadvantages for every study.

Table 2.1 Overview of the recent Arabic sentiment analysis researches

Work	**Data source**	**Data size**	**Language/ dialect**	**Approach**	**SA level**	**Advantages and Disadvantages**
Farra et al. (2010) [40]	Movie reviews	44 documents (27 positive, 12 negative and 5 neutral)	Modern Standard Arabic	LB and grammar based	Sentence-level and Document-level	(+) The uses of new grammar based method. (-) PoS not used in lexicon generation.
Rushdi-Saleh et al. (2011) [27]	Movie reviews	500 (250 positive and 250 negative)	Modern Standard Arabic	SVM and NB supervised	Document-level	(+) Presenting OCA Arabic sentiment corpus. (-) Small size corpus. (-) Neutral category is not considered.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
Shoukry and Rafea (2012)[47]	Twitter	1,000 tweets (500 positive and 500 negative)	Modern Standard Arabic / Egyptian dialect	SVM and NB supervised	Sentence-level	(+) Identification and adding of ineffective words for the Egyptian dialect. (-) Corpus size is small. (-) The neutral category is not considered.
Abdul-Mageed et al. (2012) [58]	Social networks: (chats, Twitter, forums,blogs and Wikipedia)	-	Modern Standard Arabic	SVM supervised	Sentence-level	(+) Introduction of morphological features. (-) No domain-specific dictionary.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
Mountassir et al. (2012) [48]	Al-jazeerah news web site	2,925 reviews	Modern Standard Arabic and English	SVM, NB and KNN supervised	Document-level	(+) Addressing the imbalances in the dataset. (-) The lack of technical experiments and results.
Elarnaoty et al. (2012) [53]	News articles	1 MB of news documents	Modern Standard Arabic	LB	Document-level	(+) The lexicon built and presented as open source to public. (-) Focus on news articles only.
Ahmed et al. (2013) [49]	Twitter	1,000 Tweets (positive, negative and neutral)	Modern Standard Arabic	SVM, BayesNet and J48 supervised	-	(+) Presentation of the challenges and solutions for Arabic SA. (-) Small-scale corpus.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
El-Beltagy and Ali (2013) [54]	Twitter	500 tweets	Modern Standard Arabic / Egyptian dialect	LB	Sentence-level	(+) Discussing difficulties for Arabic SA. (-) Small dataset size.
Al-Kabi et al. (2013a) [55]	Social media and news web sites	1.080 reviews	Modern Standard Arabic	Domain dictionary	Sentence-level	(+) Creating a domain-specific lexicon. (-) Small dataset size.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
Abdulla et al. (2013) [41]	Twitter	2,000 tweets	Modern Standard Arabic / Jordanian dialect	(KNN, D-tree,SVM, and NB) + LB	Sentence-level	(+) Arabic sentiment dictionary is built and presented as open source to public. (+) Using a combined ML and LB method. (-) Corpus size is small. (-) The neutral category is not considered.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
Badaro et al. (2014) [59]	Arabic WordNet and English Sentiment WordNet	157969 words	Modern Standard Arabic	SVM and LB	Sentence-level	(+) building Arabic version of SentiWordNet. (-) Lemma count is low and most of the terms used in social networks are not included.
Duwairi et al. (2014)[56]	Twitter	350,000 tweets	Modern Standard Arabic / Arabizi+ Emoticons	SVM, NB and KNN supervised	Sentence-level	(+) Presented a frame provides SA of Arabic dialects, Arabiz and expressions. (-) The lexicons and the dictionaries need to be expanded.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
El-Makky et al. (2015)[52]	Twitter	-	Egyptian dialect	SVM supervised	Sentence-level	(+) Create a new lexicon by combining two Arabic sentiment lexicons (MPQA and ArabSenti). (+) Using the Semantic Orientation algorithm for SA. (-) Corpus size is small. (-) The neutral category is not considered.

Work	Data source	Data size	Language/ dialect	Approach	SA level	Advantages and Disadvantages
Duwairi et al. (2016) [57]	Twitter	3206 tweets	Arabizi	SVM and NB supervised	Sentence-level	(+) Creating a dataset for Arabizi SA. (-) Neutral class weaker than negative class in the dataset.

2.4. GDELT

There are many incidents happening throughout the world in the last 24 hours and that are worthy of being news in the mainstream media. These events which are captured and updated every 15 minutes from 1979 to present by GDELT "(Global Database of Events, Language, and Tone)" project, can only be defined as a big data. GDELT put all these data at the disposal of all researchers worldwide as open-source big data [33]. Every 15 minutes GDELT is scanning the world's mainstream news media, as well as the social media, multimedia objects, and the environment of digital library characteristics such as DTIC, JSTOR to obtain GDELT codified metadata. This annotated metadata stored and indexed in GDELT databases [60].

If the language of the scanned source text is one of 65 different languages other than English, GDELT source language identifier is triggered. Currently, for 50 languages out of 15 languages (Arabic, Basque, Catalan, Chinese, French, Galician, German, Hindi, Indonesian, Korean, Pashto, Portuguese, Russian, Spanish, Urdu), news text is depicted to English in real time and then the natural language processing mechanisms are engaged to record the inferred assets and the tags and metrics for each entity in the database. These 15 languages are directly passed directly into the analysis process without the translating into the English language through existing dictionary sub-structures, thus allowing analysis without loss and incoherency due to translation [61].

In general, it's seen that the requirements which forms the basic pillars of the concept of big data and included in the literature as 5v [62], (Volume, Velocity, Variety, Veracity, Value) are found in the GDELT. Also, by being an open source of big data, GDELT will be used as a basic data source for the academic world for decision support processes in the near future so that the researchers, executive powers, conjuncture-based decision-making and investment specialists will be able to capture the moments in the world.

GDELT presents essentially two main datasets: "Events" and "Global Knowledge Graph (GKG)". These datasets use "Conflict and Mediation Event Observations" (CAMEO) [63] coding for recording events and saved in CSV file format.

The GKG Database keeps track of people, organizations, companies, positional data, and the data tagged with theme and sentiment tags, from each news source scanned.

In our study, we used GKG Dataset to obtain URLs of the news and their tone values. The tone value between +100 and -100 can represents the sentiment score related to a specific news article.

To interact with the databases and datasets that offered by GDELT, Google Big Query is used together with Structure Query Language (SQL). The data obtained from GDELT databases can be accessed via Google's Cloud Storage and downloaded in forms of CSV files [64].

2.5. Machine Learning Approaches for Sentiment Analysis

ML-based approaches have been widely used and preferred for SA application for a long time due to its performance and reliability. In this approach, a pretrained ML decision model is used to identify the sentiment polarity of the target text. This ML-based decision model can be built via training an ML algorithm using a sentiment dataset which can be optioned form a sentiment specific corpus. There are two types of ML algorithms used to solve two different problems, one for classification problem and the other regression. SA can be considered as an ML classification problem when the target text can be classified into one sentiment category among different categories such as positive, negative and neutral. SA also can be considered as ML regression problem, when the ML model can predict a numerical value that represents the sentiment strength score of the target text [65].

2.5.1. Machine Learning Approaches for Sentiment Classification

SA can be considered as classification problem when an ML-based classifier tries to assign the input text into a predefined category such as (positive or negative). Machine learning classification method used for assigning an unknown instance to a specific class label based on a classification model built using a set of instances with known class labels [65] [66].

2.5.1.1. Support Vector Machine (SVM)

In support vector machines classification method, the kernel that is shown as training set vectors is addressed in a space with higher dimension by using the kernel function [15]. These kernels have types such as linear, polynomial, Radial Basis Function (RBF) and sigmoid. The use of the function in DVM modeling was determined by Vapnik and Cortes (1995) [67] as follows:

$$y_i(W^T \emptyset(x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \tag{2.1}$$

Provided that,

$$\min_{W,b,\xi} \frac{1}{2} W^T W + C \sum_{i=1}^{l} \xi_i \tag{2.2}$$

While support vector machines are used, accurate modeling of the data and determination of the correct parameters with the correct kernel are very effective for the success of the model. Therefore, before using SVM; the dataset should be measured between the range of [0,1] or [1,1] if possible and the experiments should be made with verification sets until the best parameters are obtained. Failure to measure the data in a correct manner may lead to the absence of outcomes for SVM classifier and also a failure to select the correct parameters may lead to poor model performance.

2.5.1.2. Naive Bayes Algorithm (NB)

Naive Bayes is a classification method that developed based on the Bayes probabilistic theorem. It’s an approach that calculates the likelihood of a new data belonging to any of the existing classes by means of using the example data in the presently classified case. In this classifier, qualifications are independent of each other. All the samples have the same level of significance. The value of a feature does not contain information about the value of another feature.

Let's imagine that we are working on a set of data, each consisting of *n* qualities and included in any of the *m* classes. If we want to classify a new sample of X whose class is unknown in this case, the probability of the sample belonging to that class is calculated for each class by means of using Equation (Eq 2.3). The class with the highest probability among these values is regarded as the class to which the sample belongs.

$$P(S_i|X) = \frac{P(X|S_i) * P(S_i)}{P(X)} \tag{2.3}$$

$P(S_i|X)$ probability of occurrence of Si event when X event occurs,

$P(X|S_i)$ probability of occurrence of X event when Si event occurs,

$P(S_i), P(X)$ the prior probability of Si and X events.

The value of P (X) is the same for each sample data since each X sample has the same rank significance. In this case, Equation (Eq 2.3) can be simplified to Equation (Eq 2.4).

$$P(S_i|X) = P(X|S_i) * P(S_i) \tag{2.4}$$

For each class, the class to which the sample belongs is found after Equation (Eq 2.4) is applied and the probabilities are calculated [68].

2.5.1.3. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is a robust stochastic method for classifying observed data samples of discrete-time series [69]. There are three parameters to be estimated in hidden Markov model. One of them is the state transit probability matrix A, which contains the a_{ij} elements that indicate the likelihood of passage of state at the moment *i* and in the case of t +1 j.

When an observation sequence with O = $\{o_1, o_2, \ldots, o_T\}$ is defined, each element of this vector sequence indicates the feature vectors used in the classification systems. B = $[b_j(o_t)]$ is observation symbol probability distribution matrix. It indicates the observation probability of $b_j(o_t)$, o_t vector at *t* moment and *j* case. The vector π= {π_i} indicates the initial state distribution stating the probability of being in the state of *i* at the beginning. These three parameters form the Hidden Markov Model with λ = {A, B,π}. Apart from these, the number of states is N, the number of hybrids is M in each case. There are also various methods to show the probability of observance symbol, but Continuous Probability Density Function is the one that is preferred most among these methods. HMM can be used to perform classification task by training separate HMM for each class and then the model that has the highest likelihood is selected. Finally, the classification decision is done by assigning to the class that maximizes the posterior probability [69].

2.5.1.4. K-Nearest Neighbor Algorithm (KNN)

K-Nearest Neighbor algorithm (KNN) is considered as one of the simplest pattern recognition methods that classify an unknown instance based on the class of the closest training instances in feature space. This algorithm makes the class classification process according to the class of nearest neighbor as the provided *k* value. Classification of a vector in the kNN algorithm is performed using vectors *n* whose classes are known. The sample to be tested is processed individually with each sample in the training cluster. The *k* which is nearest to sample in the training cluster is selected in order to determine the class to be tested. It's concluded that the sample to be tested belongs to whichever class has the most samples within the cluster consisting of selected samples. The distances between the samples are found by means of *Euclidean* distance. Equation (Eq 2.5) is the Euclidean distance formula giving the distance between 2 n-dimensional points [70].

$$d(x,y) = \sqrt{\left\{\sum_{i}^{n} (X_i - Y_i)^2\right\}} \qquad (2.5)$$

2.5.1.5. Logistic Regression (LR)

Logistic regression technique is based on the concepts of probability and odds ratios [71]. Odds are the ratio of the number of results of a given type to the total number of occurrences. In the logistic regression, the odds ratio is defined as the probability of non-occurrence for the occurrence possibility of an event. In other words

Odds ratio:

$$\frac{p}{1-p} \tag{2.6}$$

If the probability of occurrence for a " p " event (success factor) in this case is 1- p, this shows the non-occurrence possibility of the said event (realization of the failure factors).

The odds ratio can be greater than 1, smaller, and equal to 1, depending on the ratio.

Logistic regression analysis is considered as non-linear analysis. The key concept in the logistic regression is the "logit" concept. Logit is the logarithm of odds ratios. Starting from this point, the logistic regression model to be estimated can be shown as:

$$logit(p) = log\frac{p}{1-p} = x'\beta + u \tag{2.7}$$

Here, p is the realization ratio of the case determined as success factor for the dependent variable; xk is the number of the independent variables that involved in the independent variable matrix with dimension of $n \times (k+1)$; $\beta \times (k + 1)$ represents the parameter vector and u is the error term. With the help of odds ratio, the success factor of each dependent variable on probability can be obtained by the equation (Eq 2.8):

$$p = \frac{\exp(x'\beta)}{(1+\exp(x'\beta))}. \tag{2.8}$$

The obtained probability value p form transforming the logit function for a certain input is then mapped to two or more discrete classes to achieve the classification task.

2.5.2. Machine Learning Approaches for Sentiment Regression

SA can be also considered as regression problem when an ML-based regression approach used for estimating the sentiment score of the input text based on learned function. ML regression approaches used to build a model by fitting a function f(X) which can describe the best correlation between input X and continuous real value output Y then use this learned model to predict the real value output for a specific input [72] [65].

2.5.2.1. Multiple Linear Regression (MLR)

the objective of multi-linear regression is estimating the dependent variable value based on the independent variables that affect the dependent variable and to find out which of the independent variables has more affects the dependent variable more.

In multiple regression; if the independent variables are $x1, x2, \ldots., xp$ and dependent variable is y, the relationship between them is expressed by Equation (Eq 2.9).

$$y = b_0 + b_1x_1 + \cdots + b_jx_j + b_px_p + \varepsilon \tag{2.9}$$

Here; $b_{0,}$ $b_{1,}$ $b_{2,}$ ..., b_j ... b_p are called the regression coefficients of the unknown. When other variables are kept constant (when the effect of other variables are eliminated), any regression coefficient of b_j represents the amount of expected change in the variable y in return for a one-unit change in x_j variable. In other words; $b_{1,}$ $b_{2,}$ ..., b_j ... $b_{p;}$ are the relative contributions of the independent variables to the determination of y. Thus, b_j (j = 1, 2, ..., p) parameters are often referred to as partial regression coefficients. b_0 is called as the cut-off point or constant, and it represents the value of dependent variable when all x_j variable values are zero. "ε" is the error term [73].

2.5.2.2. Support Vector Regression (SVR)

Support Vector Machine (SVM) Is a kernel-based method which can be used for solving both of ML classification and regression problems. This learning strategy was developed by Vapnik [16] and is a very robust method based on principles in ML algorithms.

Support Vector Regression method is aiming to find the function with the lowest generalization error. The general expression of regression with support vector machines is as following:

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{2.10}$$

Where $\alpha_i,\ \alpha_i^*$ and b are Lagrange coefficients and the regression is calculated to minimize the risk function. $k(x_i, x)$ is the kernel function. In support vector machines, generally linear, polynomial, sigmoid and radial basis kernel functions are used.

2.5.2.3. Multilayer Perceptron (MLP)

This method first appeared in the literature by McCullough and Pitts, who proposed the cell model in 1943 [74]. The ability of the brain to perform difficult operations and comprehend the complex samples, and especially the ability to learn only some of the essence without knowing the physical relationships involved, inspired scientists to develop the Artificial Neural Networks ANN method. Artificial neural networks can be regarded as a black box producing output in response to inputs.

The basic logic of artificial neural networks is relied behind the identification of the weight coefficients between the input and output of the problem and constructing this process with a learning system for each input-output from the point of biological nerve cell structure. Artificial neural networks are dense parallel systems consisting of many processing elements connected by different weights. Among the most common methods used in ANN methods are those based on the principle of backpropagation.

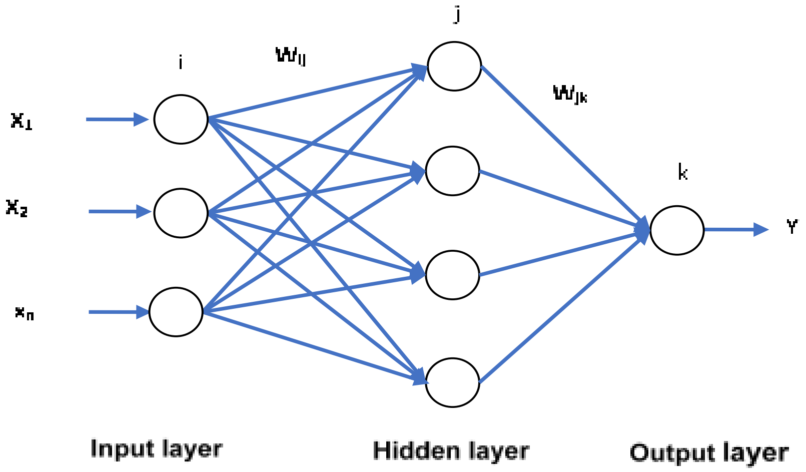


Figure 2.1 A Multilayer Artificial Neural Network architecture

Multilayer Artificial Neural Networks mainly consists of three layers as is illustrated in Figure 2.1. These layers are called: the input, hidden and output layers respectively. ANN and MLP are considered as a powerful ML tools which can be used for both regression and classification [73].

2.5.3. Machine Learning Based Model Ensemble Techniques

In ensemble learning, multiple ML-based models are cooperatively works together for solving the same problem. An ensemble classifier combines the decisions of the individual weak classifiers and aims to enhance the accuracy final decision and produce a stronger classifier. There are basically two approaches for combining classifiers, one approach is to use similar classifiers and to combine them together using techniques such as Bagging, Boosting or Random Subset. A second approach is to combine different classifiers using model fusion using Stacking technique [75].

2.5.3.1. Bagging

In this method, different training sets are used for training multiple classifier models from the same type. A method based on sampling and replacement is applied for creating the multiple training sets that used in bagging method. The decision of classifying an unknown instance is done with respect to the majority voting of all results that obtained by the ensemble classifier models [76].

2.5.3.2. Boosting

In this method, different training sets with weighted instances are used for training multiple classifier models from the same type sequentially. This method focuses on the training samples that misclassified by the previous classifiers in the chain, by using higher weights to the misclassified instance before passing it to the next classifier. The final decision is obtained by combining decisions of base classifiers by a voting scheme [76].

2.5.3.3. Random Subspace

This method is similar to bagging, but the difference that it's selects a random subset of features from the dataset instead of instances. In random subspace, different training sets with different features subspaces are used for training multiple classifier models from the same type. If there are many of irrelevant and redundant features in training dataset, so using random subspaces may results in overcoming these unwanted features, since it creates multiple training sets with different features subspaces drawn randomly from the original dataset. Similar to the other ensemble methods the final decision is obtained by combining decisions of base classifiers by a voting scheme [76].

2.5.3.4. Stacking

Staking is a technique that fuses multiple classifiers applied to a specific classification problem and aims to improve the results of the individual classifier [77].

Staking method combines multiple classifier models from different types using another classifier called meta-classifier in a stacked structure. The meta-classifier is trained based on the output of each combined model using staking.

The classification task in Stacking is achieved in two stages. In the first stage each one of combined model generates the classification decision for the unknown instance, then is the second stage these output decisions are fed as input to the meta-classifier which is, in turn, provides the final classification decision for the unknown instance [77] [78].

2.6. Concept-Based Sentiment Analysis

Concept-based SA methods are superior to standard word-level SA methods because they take into account the meanings of multiple word expressions. Concept-based SA approaches are concentrate on the semantic analysis of the textual contents through using semantic networks such as (SenticNet) and web ontologies, in order to extract the concepts that associated with the natural language opinions [35] [38].

Concept-based emotion analysis is taking steps away from methods that use blind keyword and word co-occurrence frequencies, based on ontologies or semantic networks. The concept-based emotional analysis provides a better understanding of texts and offers a significant enhancement in the performance of the model [34]. Concept-based approaches can also detects complex emotions [38].

The first step to Concept analysis was made by Wille in 1982 [79] when he presented a “Formal Concept Analysis” (FCA) which is a mathematical model used for analyzing and visualization data (configuration, analysis, and visualization) and it’s based on the concept of duality known as Galois connection [80]. Formal concepts are considered as formal summaries which involve clusters of data assets and their properties. Conceptual patterns are the type of conceptual structures which are consist of objects with their attributes that belong to specific areas. They are formed by specifying the objects and then their relations are demonstrated. The Fuzzy Formal Concept Analysis (FFCA) approach presented in [81] showed a great success in addressing the uncertainty information issues.

In [81] an FFCA based classification framework is proposed to classify documents based on its conceptual summaries. The classification model is trained based on concepts using FFCA method. Thus, they intended to reduce the uncertainties that effect the classifier performance. They have studied the polarity datasets of benchmark test bed (Reuters 21578) and two views on film and eBook interpretations. They have achieved good results in all data sets and have proved that the noisy drop sensitivity ability is good.

The work presented by Kontopoulos et al. [82] have adopted the FCA approaches for constructing an ontology field model. They used an ontology-based technique from their Twitter posts to make a more effective SA by dividing each tweet into view sets tailored to the topic. They have worked on Smartphone spaces. The architectural views they use give a more detailed analysis of their posts. This also makes it possible to distinguish the specific characteristics of the subject from the scores given to the subjects.

One of recently developed concept-based SA approaches is called pSenti and presented in Mudinas et al. [83]. This system integrates the learning-based approaches with the data dictionary based Opinion Mining (OM). The authors claim that the pSenti system has acquired a high emotion polarity classification performance in term of accuracy. At the same time, pure data dictionary has been compared with base systems in order to find emotion strength. They have tested the pSenti system using IMDB movie reviews and CNET software reviews datasets and they showed that pSenti has performed better than most current system-like hybrid approaches such as SentiStrenght.

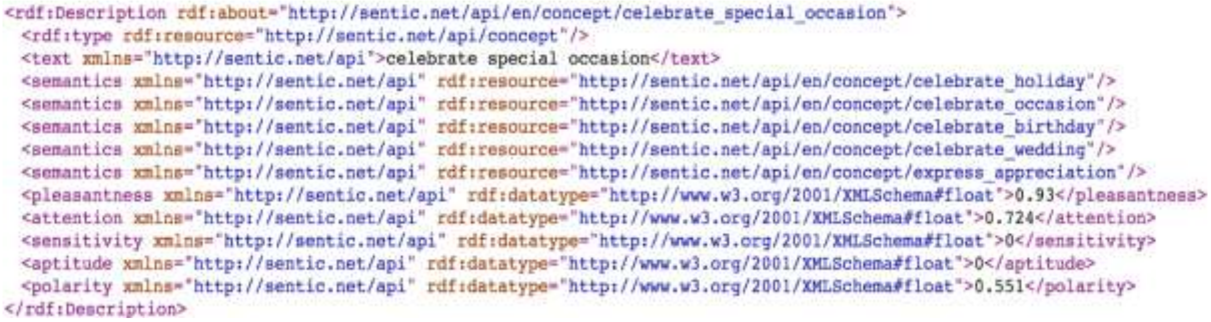
Cambria et al. [84] have introduced SenticNet. They have developed SenticNet which act as a semantical link between concept-level emotion and natural word-level language data. They have built their systems with Sentic computation which is an integrated framework that taking advantage of SemanticWeb and Artificial Intelligence (AI).

2.6.1. SenticNet

SenticNet is a concept-based sentiment lexicon that can be considered as one of the important resources that can be used for building a concept based SA system. SenticNet use graph mining and multidimensional scaling to reduce the gap between the word and the opinions that covered by the words in natural language. Many applications have been developed by employing senticNet. These applications can be exploited in many fields such as the analysis of a considerable amount of social data, human and computer interactions [84]. SenticNet_v3 consist of a 30k single and multi-word concepts while SenticNet_v4 contains 50k of concepts. SenticNet provides different information about each concept, this information includes [85];

- Polarity which is a float number in the range between -1 to 1 that represents the sentiment score of the input concept.
- Five different single or multi-word senses that semantically related to the input concept.
- Four different values that represent the diminutions of the hourglass emotion for the input concept.

An example of the SenticNet contents is shown in Figure 2.2 below.



```xml
<rdf:Description rdf:about="http://sentic.net/api/en/concept/celebrate_special_occasion">
  <rdf:type rdf:resource="http://sentic.net/api/concept"/>
  <text xmlns="http://sentic.net/api">celebrate special occasion</text>
  <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_holiday"/>
  <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_occasion"/>
  <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_birthday"/>
  <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/celebrate_wedding"/>
  <semantics xmlns="http://sentic.net/api" rdf:resource="http://sentic.net/api/en/concept/express_appreciation"/>
  <pleasantness xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.93</pleasantness>
  <attention xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.724</attention>
  <sensitivity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</sensitivity>
  <aptitude xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</aptitude>
  <polarity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.551</polarity>
</rdf:Description>
```

Figure 2.2 A sample of the SenicNet content

SenticNet is automatically built by applying dimension reduction techniques and graph mining on different knowledge-based resources such as WordNet, ConceptNet, and SentiWordNet [84].

2.6.2. SenticNet Based Sentiment Analysis Applications

In [86] SenticNet were used for developing a system for the detection of the polarity of contextual concepts based on Bayesian approach. In the work presented in [87], the authors designed a semi-supervised approach based on fuzzy SVM classification technique which is used to determine the polarity of the SenticNet concepts. The main goal of the proposed approach was to improve and enhance the SenticNet. To train the concept based SA model, the authors used a different syntactic and lexical feature sets together with the SenticNet based features. In [88] Qazi et al., a supervised classification method is used for developing an approach for enhancing the performance of business intelligence using the SA of the suggestive reviews. The proposed system utilizes SenticNet to extract a sentiment based features as well as discovering the domain of the context. In work presented in [89], the authors introduced a method for extracting the concepts from the sentence for concept-based SA. They introduced a method that takes advantage form combining both the SenticNet and the WordNet together for concept extraction task. Araújo et al. in [90] introduced an e-health system called iFeel which aims to analyze the patient's opinions about the provided healthcare. This system utilizes both of SenticNet and SentiWordNet for the SA task. In [91] a multilingual lexicon for concept based SA is built using some approaches that are similar to the approaches used for building the SenticNet. Methods in the works that presented in [86] [92] [93] are also used SenticNet for developing varied supervised approaches. In [94] a supervised concept based SA system was built. This system is also getting benefit from SenticNet which is used for concept extraction and generate the bag-of-concepts features that in turn used for building the supervised SA model. In [95] the authors used a random walk based method on ConseptNet which aimed to extend the SenticNet by adding more concepts. They also used this extended version of the SenticNet to generate a set of features called Bag-of-Sentimental-Concepts where each concept in the features vector is represented by the TF-IDF multiplied by its polarity. Bravo-Marquez et al. [96] developed an approach that emphasizes both of SentiWordNet and SenticNet and aims to improve the performance of the supervised SA system for Twitter. In [97] an unsupervised classification system for short text message (SMS) is presented. This system also uses and utilize SenticNet for the task of assigning the sentiment polarity to each SMS. In [98] a Document-level classification system based on sentiment similarity is presented. In this system, SenticNet is adopted to identify and extract the sentiment based features of each document and use it as a distance measure to identify the similarity.

SenticNet also have played an important role in other application that used concept-based approaches for SA tasks in various fields such as SocialWeb (Troll Filtering, Social Media Marketing, Sentic Album), HCI Human Computer Intelligent (Sentic Avatar, Sentic Chat, Sentic Corner), and e-health (Crowd Validation, Sentic PROMs Patient-Reported Outcome Measures) [99].

2.7. Word Embedding (Word2voctor)

In general case, the objective of the sentiment classification is to divide the texts into positive and negative sentiments. As it’s known, natural language processing is based on the smallest linguistic units that have independent meanings which are words. There are two commonly used methods to represent words, which are distributed representation and one-hot display which is more intuitive [100]. In One-hot encoding, words are represented as a Boolean vector being equal to the length of the vocabulary. For each word, the position 1 corresponding to the word in the representative vector and the remainder is set to 0. Although One-hot display is widely used because it’s simple and relatively easy to implement, there are some shortcomings. Among the most important shortcomings, it can be shown that even though the original words in the representation space are very similar, each vector is independent.

On the other hand, Hinton suggested a novel word distribution model known as word embedding which is different from the one-hot representation [101]. The word embedding or Word2vec represents the word as a low-dimensional vector trained by the language model and allows the related or similar words to be closer in vector space. Thus, one-hot in which the feature vectors cannot reflect the dependency relationship between words can overcome the disadvantages of the representation. On the subject of language models, Bengio [102], Collobert [103], Mikolov [104], Huang [105] have proposed different language models to improve word embedding. Bengio used a language model formed by the three-layered neural network to train the feature layer. Collobert achieved embedding with a method that simplified the artificial neural network output layer and implemented in-segment labeling called entity recognition, sentence recognition, semantic role labeling, and other natural language processing tasks based on vectors. The repetitive neural network has been used by Mikolov as a language model in which document information is fully used. Huang developed the model proposed by Collobert and increased semantic word component in embedding language. At the moment, the most popular models in the field of word-embedding are the continuous bag-of-word (CBOW) and skip-gram models proposed by Mikolov in 2013. Word embedding (word2vec) provides a vector representation for each word based on it relation with other words in the context as shown in Figure 2.3.

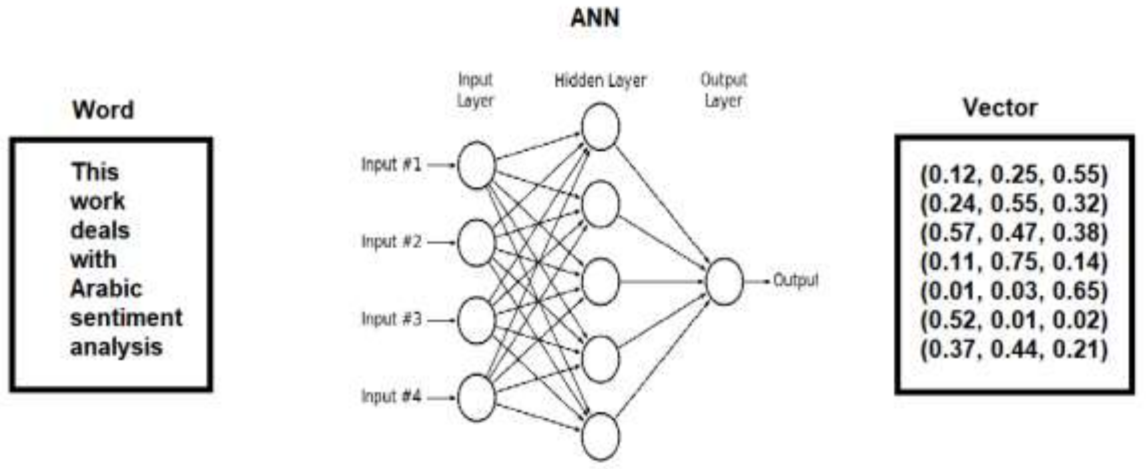


Figure 2.3 An example of word vector representation using Word2Vec

3. DATA METHODOLOGY AND MODEL

In this chapter, we present and explain our proposed approaches which, are carried out to achieve the target goals of this thesis that includes; (i) The generating of a large-scale data resources for Arabic called GLASC which stands for GDELT Large-scale Arabic Sentiment Corpus, using the metadata that provided by bigdata platform GDELT and the online Arabic news articles. (ii) The building of Document-level Arabic SA system which is based on ML approaches and utilize our GLASC corpus. (iii) The generation of concept-based sentiment lexicon for Arabic by translating the English version of SenticNet concept-based sentiment lexicon to Arabic. (iv) The building of a concept-based SA system for Arabic SA at Sentence-level based on the generated Arabic version of concept-based sentiment lexicon SenticNet and a variety of ML-based approaches.

The process of generating our GLASC corpus is explained in Section 3.1. The proposed approach for building the ML-based Document-level Arabic SA system is presented in Section 3.2. The proposed approach for translating the English version SenticNet concept-based sentiment lexicon to the Arabic language is presented in Section 3.3. finally, the proposed approach for building the Concept-based Sentence-level Arabic SA system is presented in Section 3.4. All evaluation experiment and results of our proposed approaches are represented and discussed in Chapter 4.

3.1. Approach for Generating a Large-scale Arabic Sentiment Analysis Corpus

The process of generating our (GLASC) large-scale Arabic sentiment corpus based on GDELT's metadata and the online Arabic news articles, is illustrated in Figure 3.1.

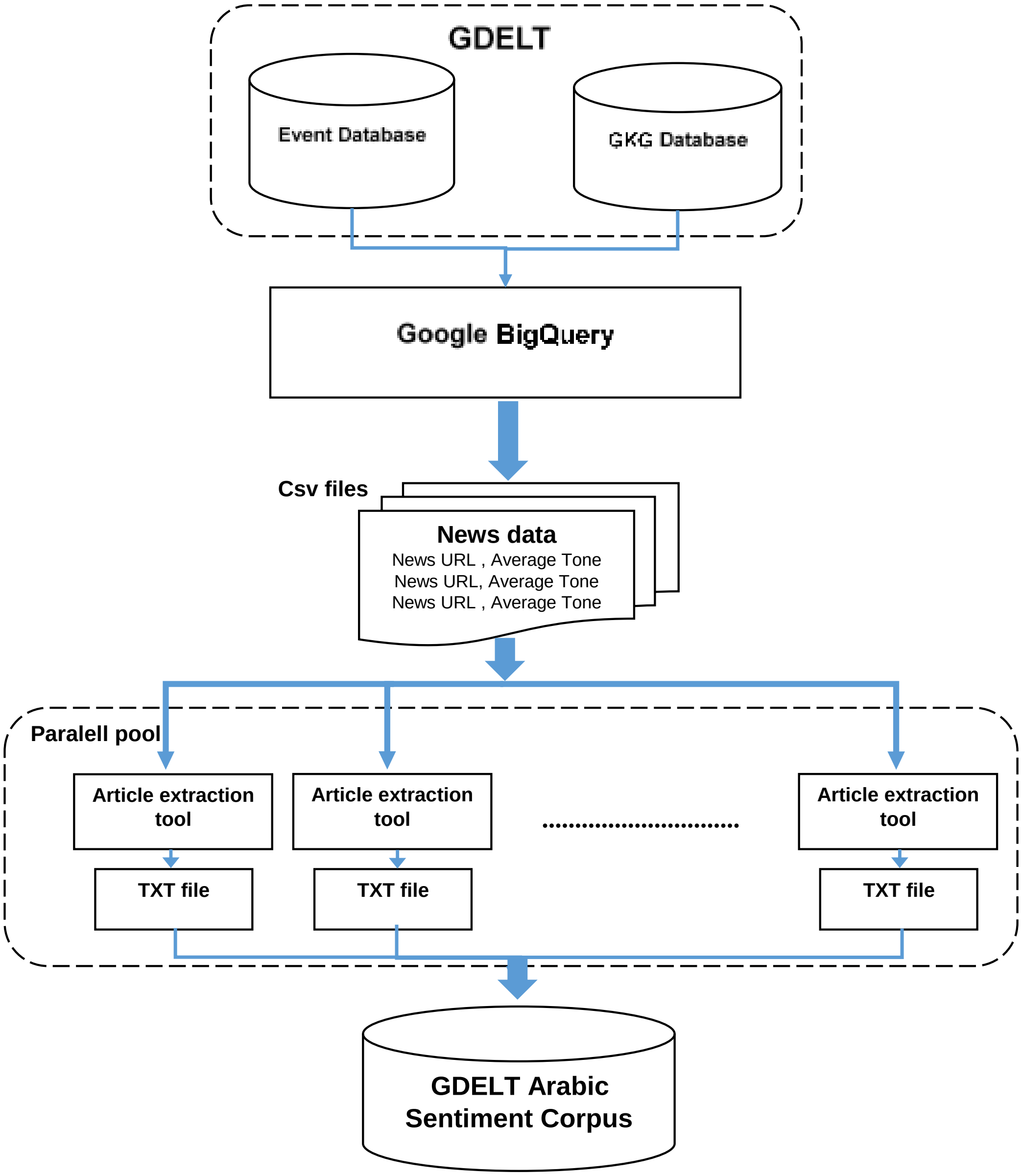


Figure 3.1 Our GLASC corpus generation process

This task of corpus generating is done in the following steps:

- Firstly, we used SQL query to fetch the data that related to Arabic news from GDELT GKG Database. GDELT stores only metadata and does not contain the news articles contents, so we can only be fetching the Arabic news URLs and the corresponding Average Tone values, from GDELT. The results of this SQL query are saved into CSV file format with two columns (news URL and Average Tone) and rows are equal to the total number of the obtained news. fetching Arabic news URL from GDELT GKG database in three categories (Positive, Negative and Neutral) is done using SQL query.

- After acquiring a sufficient number of Arabic news metadata from GDELT, the next step is to obtain the contents of this Arabic news articles from the source URLs located in the CSV file that is previously obtained. For this task, we utilized an open source article extraction tool called “Boilerpipe”. When the number of the news which is required to be extracted becomes very large, the sequential extraction method which can be executed a piece at a time becomes inefficient and can be considered as time and compute intensive. In order to address this issue, we considered using a parallel article extraction method based on parallel computing. In this method, different extraction units can share the articles extracting task from different URLs and store the extracted article into text files simultaneously as shown in Figure 3.1. In the multi-core processing environment, each one of these extraction unit processing tasks that can be assigned to different CPU cores and work independently. Since our system contains a 32 core CPU, 32 extraction units are used to share the news article extraction tasks. Since the parallel extraction method can process and extract many articles in a shorter time compared to the ordinary sequential method, that can reduce the extraction time and increase the performance. Figure 3.2 shows the time required to extract and store 100 news articles using a different number of extraction units.

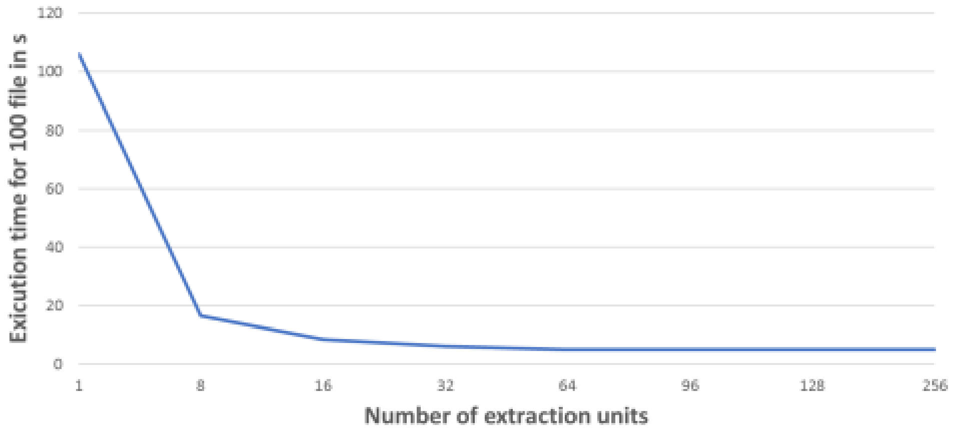


Figure 3.2 News articles extraction time with respect to the number of parallel extraction units

- The news articles contents that obtained in the previous step is stored and indexed with respects to its average tone values into three categories (Positive, Negative and neutral). When all news contents text files are indexed and assigned to the Positive, Negative or Neutral category then we applied filtering to remove the duplicated news text file and perform the final corpus.

The total number of files in our GLASC corpus which obtained from GDELT and the online Arabic news articles is shown in Table 3.1.

Table 3.1 The total number of files in our corpus

Category	Final file number after filtering	Corpus size
Negative	266,376	816MB
Positive	225,397	635MB
Neutral	218,309	448MB
Total	620,082	1.9GB

Several types of evaluation experiments will be carried out in the next chapter in order to evaluate the quality of our generated GLASC corpus.

3.2. Machine Learning Approach for Document-Level Arabic Sentiment Analysis

The architecture of our proposed approach for Document level Arabic SA using ML algorithms and our GLASC corpus is shown in Figure 3.3.

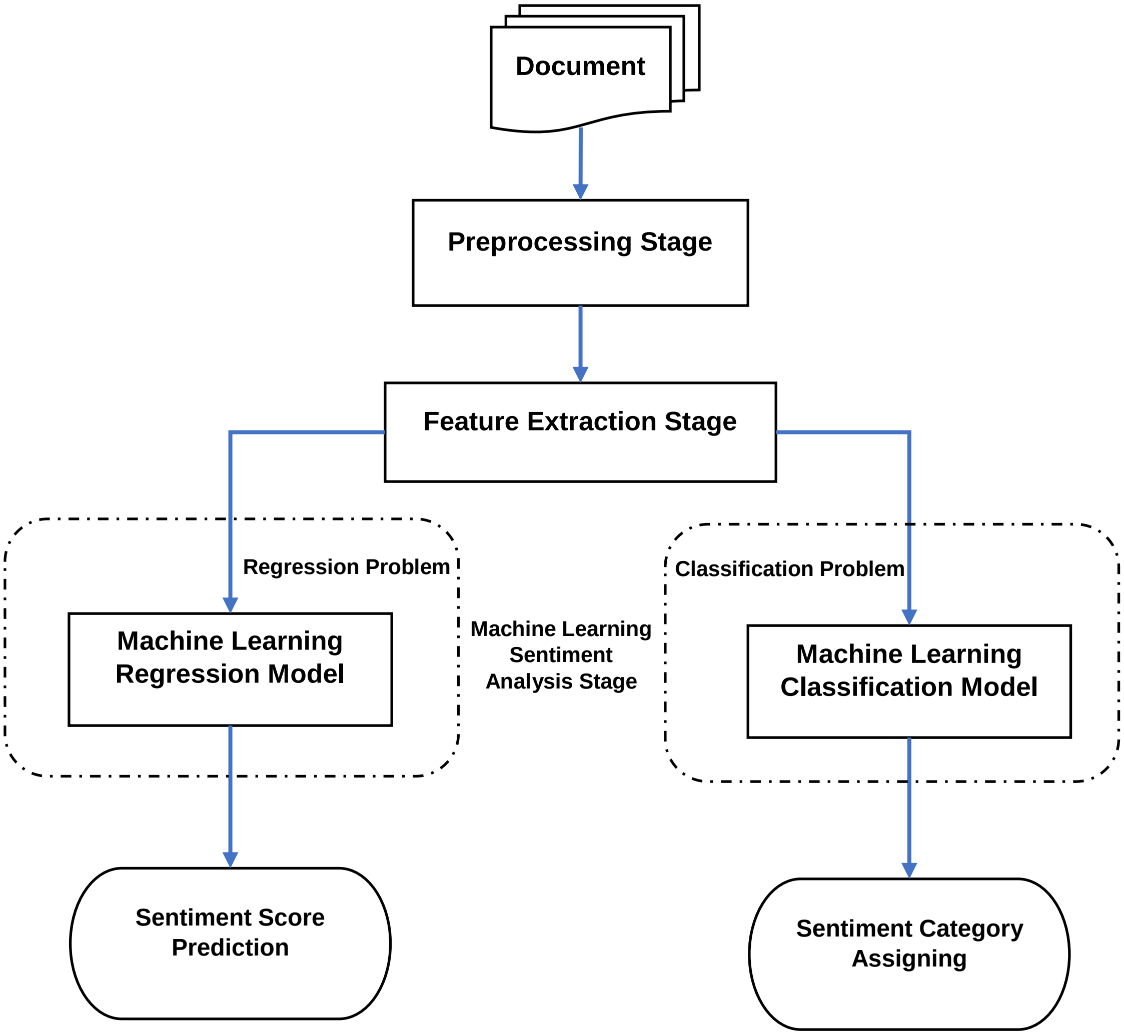

Document
Preprocessing Stage
Feature Extraction Stage
Regression Problem
Machine Learning Regression Model
Machine Learning Sentiment Analysis Stage
Classification Problem
Machine Learning Classification Model
Sentiment Score Prediction
Sentiment Category Assigning

Figure 3.3 The architecture of the proposed ML approach for Document-level Arabic SA system

The proposed architecture for our Document-level Arabic SA system consists of different stages such as (Preprocessing, Feature extraction and the ML for SA) which is explained as following;

3.2.1. Preprocessing Stage

In this stage different text preprocessing techniques are applied on the document. These techniques include, tokenizing the term in the document, removing the stop words and stemming the root of each term. For the root extraction task, we used the Buckwalter morphological analyzer's called Ara-morph Arabic lemmatizer [106].

3.2.2. Feature Extraction Stage

Since ML approaches cannot deal with non-numerical data such as raw text, this data must be transformed into a numerical form that can be understood by the ML algorithm. This transformation called feature extraction. The first stage in feature extraction of textual data is called "Bag of Words" (BoW) representation where each of the documents can be represented as a vector of terms or grams [66]. In this work, we considered using two different feature extraction methods based on Unigrams and Bigrams, to generate the features vectors. In Unigram method, each word or term in a document can be represented as a single feature, wherein Bigram method every two adjacent words can be represented as a single feature. The second stage called feature weighting and is responsible for assigning the numerical weight value to each feature in the feature vector that represents a document.

Each feature (term, Unigram or Bigram) in the vector is typically weighted using the "Term Frequency-Inverse Document Frequency" (TF-IDF) or "Term Frequency" (TF) methods [66].

The TF score of a term is a value that indicates the frequency at which the term crosses the document. While there are many terms often found in many documents that are not trivial in terms of discretization, it would be wrong to use these metrics in scoring. For this reason, IDF scores are derived. Here, the TF and IDF score for a specific term is calculated as:

$$TF(i,j) = \frac{\text{Number of times the term i appears in a document j}}{\text{Total number of terms in the document j}} \quad (3.1)$$

$$IDF(i) = \log\left(\frac{\text{Total number of documents in the corpus}}{\text{Total number of document contain term i}}\right) \quad (3.2)$$

$$TF - IDF(i,j) = TF(i,j) \times IDF(j) \quad (3.3)$$

3.2.3. Machine Learning based Sentiment Analysis Stage

In this stage, we considered the SA task as two distinct ML problems which are classification and regression problems. For the classification problem, an ML classification model is used to classify and assign the input document into one of three different categories (Positive, Negative and Neutral) depends on the sentiment orientation of this input document. The ML classification model can be built by training an ML classifier algorithm on a dataset. In this work, we considered using five different ML based classification algorithms which are widely used in SA such as (SVM [15], HMM [69], NB [107], ANN [108] and KNN [109]), to build the classification model for the Arabic document level SA. In the addition to that, we also considering investigating a several types of ensemble learning methods such as (Bagging, Boosting, Random Subspace [77] and classifier model fusion using staking method [78]) in order to verify its effect of improving the classification performance for SA.

In the other hand, for the regression problem, an ML regression model tries to predict the sentiment score of the input document in the term of real value in the range [-1 to 1] depend on the sentiment strength of this input document. In this work, we adopt different types of ML regression algorithms to build the regression model such as (SVR [110], MLR [111] and MLP [112]), which is responsible for the prediction of sentiment score for an Arabic document.

These classification and regression algorithms can be trained using different datasets which can be generated from our Arabic large-scale SA corpus GLASC. The processes of generation these datasets will be presented in the next chapter together with our evaluation process for the proposed sentiment classification and regression models.

3.3. Generation of the Conceptual Based Arabic Sentiment Lexicon (Ar-SenticNet)

The task of generating the Arabic version of SenticNet [85] concept-based sentiment lexicon is consisting of two stages as shown in Figure 3.4. The first stage involves the process of translating each concept found in the English version of SenticNet and the second stage is involving the extension process of the translated Arabic version of the SenticNet.

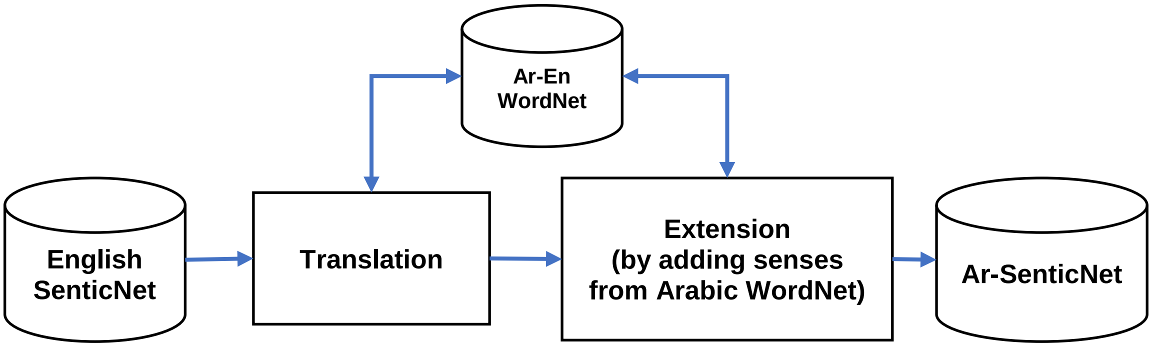


Figure 3.4 Arabic SenticNet generation process

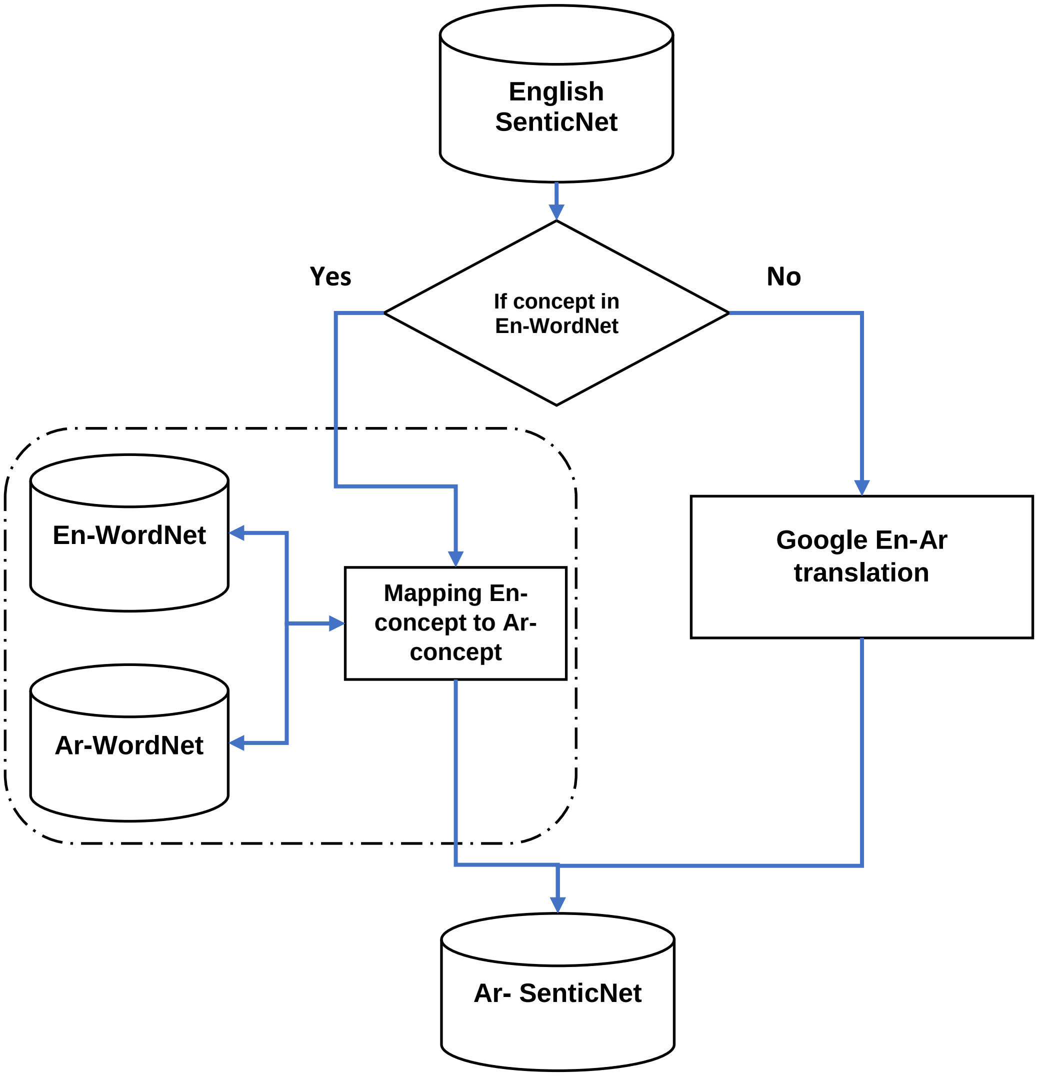


Figure 3.5 The proposed approach for translating English SenticNet to the Arabic language

3.3.1. The Translation Process

The translation process of the SenticNet conceptual lexicon to the Arabic language that shown in Figure 3.5 is done in two phases. WordNet is considered as one of the resources that used to build the SenticNet, so that in the first place we used a cross-language translation to translate SenticNet concepts to Arabic based on the mapping of both English WordNet (En-WordNet) [36] and Arabic WordNet (Ar-WordNet) [113]. The English concept that is required to be translated into Arabic is firstly searched in English WordNet and If it’s found in the English WordNet, then a mapping between

English and Arabic WordNet is used to obtain the Arabic translation of this concept. The second phase, in case, that the concept that required to be translated into Arabic is not found in the English wordnet, then the concept is translated into Arabic using Google Translation API. Some examples of SenticNet concept translation to Arabic is shown below;

- SenticNet--> trip_up ---->En-WordNet (found)--- mapping --- Ar-WordNet ---> أَمْسَكَ
- SenticNet--> switch_off ----> En-WordNet (found)--- mapping --- Ar-WordNet ---> أَوْقَفَ
- SenticNet--> religious_ceremony ----> En-WordNet (found)--- mapping --- Ar-WordNet ---> اِحْتِفَال دِينِي
- SenticNet--> care ----> En-WordNet (found)--- mapping --- Ar-WordNet ---> اِهْتَمَّ

- SenticNet--> catch_fire ----> En-WordNet (not found) ---- Google translate ----> اشتعل
- SenticNet--> change_hair_style ----> En-WordNet (not found) ---- Google translate ----> تغيير تسريحه الشعر
- SenticNet--> long_trip ----> En-WordNet (not found) ---- Google translate ----> رحلة طويلة

3.3.2. The Extension Process:

Each concept with SenticNet has a set of senses called semantics. The aim of the extension process is to extend this set of senses by adding more senses which can be obtained from WordNet. The translated version of SenticNet is extended as following; (i) Searching all concepts that are translated into the Arabic in Ar-WordNet, (ii) Obtaining the synonym sets for each concept that found in Ar-WordNet. (iii) Adding these synonyms sets to Ar-SenticNet to extend the senses set for the concepts. An example of the concept extension by adding sense from Ar-WordNet, is shown below.

- "قائمة الحساب"----->Ar-WordNet---->(found)------>get synset---->

  [حسبة,تعداد,احتساب,فاتورة الحساب,حساب,كشف,تقدير,عد,عداد]

We applied this translation and extension approach on both English SenticNet_v3 and the recently released SenticNet_v4 to generate two different versions of Arabic SenticNet.

SenticNet_v3 is consist of 33k English concepts. After applying our translation approach described in Section 3.3.1, 7k English concepts were translated into Arabic using WordNet mapping method and, another 24k of English concepts were translated into Arabic using Google translation. This resulted in a 31k of Arabic concepts. Then by applying the extension process, 10k of the translated concepts were found in Ar-WordNet and, the obtained synonyms set for these concepts are added another 7k sense to extend the number of concepts in Arabic SenticNet. This resulted in a total of 38k of Arabic concepts and thus the Ar-SenticNet_v1 is generated.

We also used a similar approach to translate the SenticNet_v4, which is consist of 50k English concepts. 9.4k of SenticNet_v4 concepts were translated into Arabic using WordNet mapping and, by using the Google translate method another 30k of concepts were translated. This resulted in a 39.4k of Arabic concepts. Then by applying the extension process, 11.2k of the translated concepts were found in Ar-WordNet and, the obtained synonyms set for these concepts are added another 9k sense to extend the number of concepts in Arabic SenticNet. This resulted in a total of 48k of Arabic concepts and thus the Ar-SenticNet_v2 is generated. Figure 3.6 shows a comparison between the generation process of our two concept-based sentiment lexicons Ar-SenticNet_v1 and Ar-SenticNet_v2.

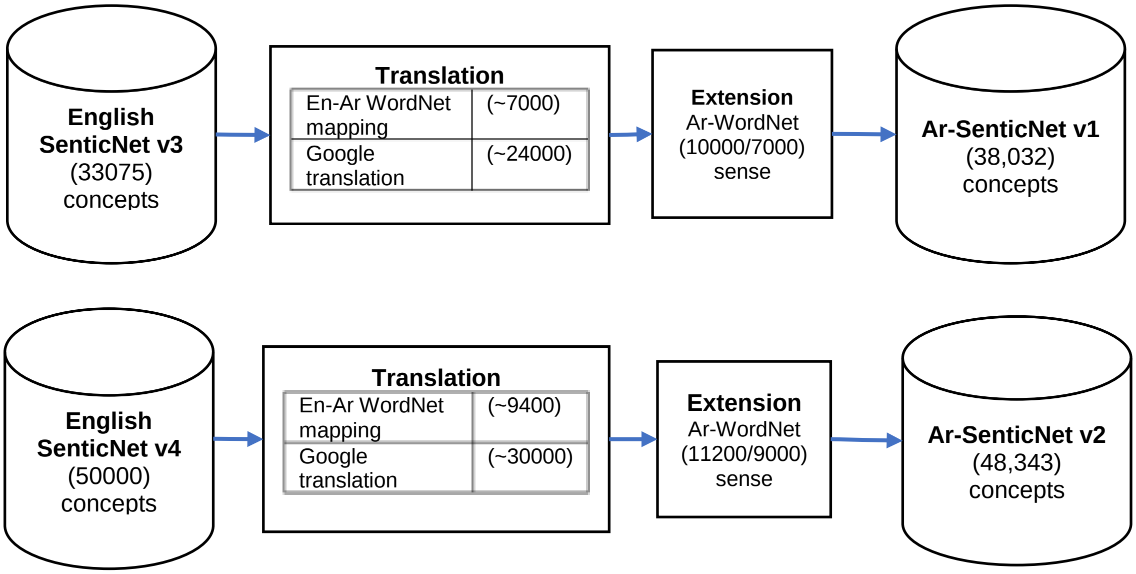


Figure 3.6 A comparison between the two concept-based sentiment lexicons Ar-SenticNet_v1 and Ar-SenticNet_v2 generation process

3.4. Concept-based Approach for Arabic Sentence-Level Sentiment Analysis

The proposed architecture of our ML approach for Sentence-level concept-based Arabic SA system is shown in Figure 3.7.

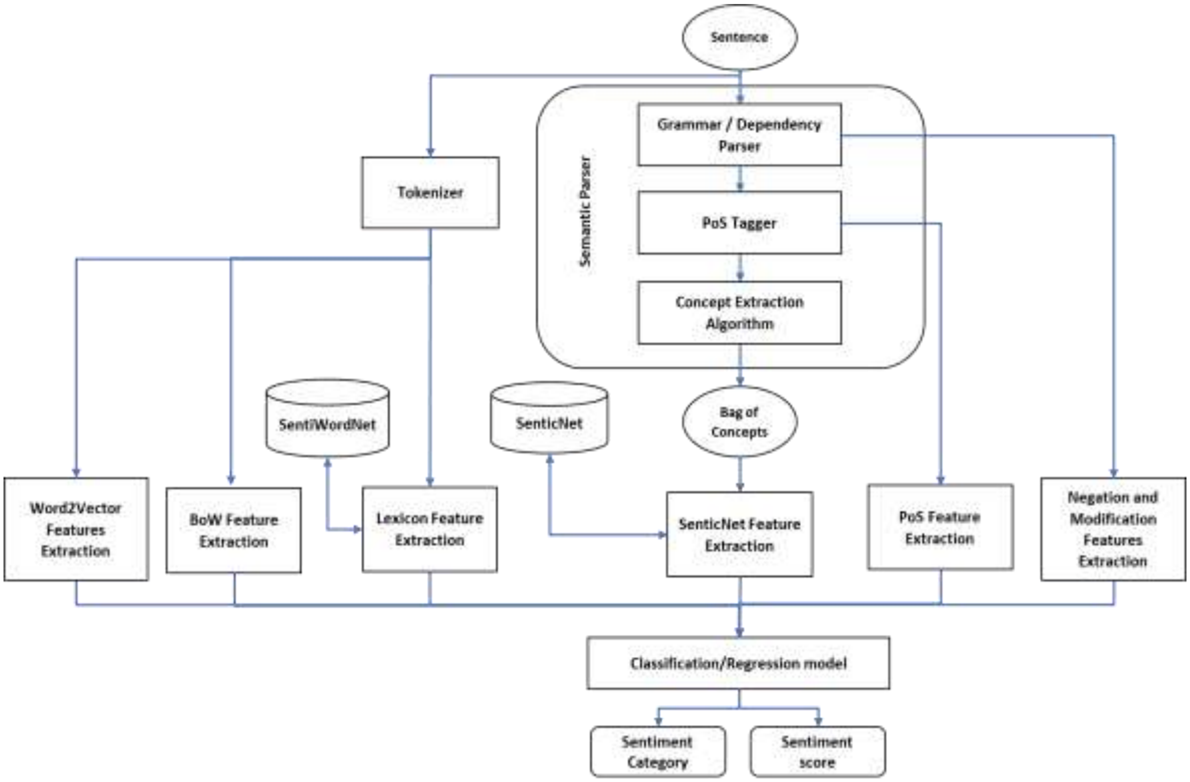

Sentence
Semantic Parser
Grammar / Dependency Parser
PoS Tagger
Concept Extraction Algorithm
Tokenizer
SentiWordNet
SenticNet
Bag of Concepts
Word2Vector Features Extraction
BoW Feature Extraction
Lexicon Feature Extraction
SenticNet Feature Extraction
PoS Feature Extraction
Negation and Modification Features Extraction
Classification/Regression model
Sentiment Category
Sentiment score

Figure 3.7 The architecture of the proposed ML approach for Sentence-level concept-based Arabic SA system

The architecture of our concept-based Arabic Sentence-level SA approach shown in Figure 3.7 involves different stages such as (Concept extraction, Feature extraction and the ML algorithms for the SA task). These stages are explained as follows;

3.4.1. Semantic Parser (SP)

The concepts can be simply defined as the single or multi-word expression that carries the meaning of a phrase or sentence. Semantic parsing is considered as the task of extraction the concepts for the sentence based on its grammatical structure. Concept extraction process involves fragmenting and partitioning the sentence into a noun and verb clauses then form a candidate list of words that match the grammatical rules of the concept in those parts.

Step 1. "Stanford Arabic parser" [114] is used to extract noun and verb phrases from the sentence.

Step2. “Stanford Arabic Tagger” tool [114] is used to assign the part of speech tags to each word found in the noun and verb phrases. Figure 3.8 shows an example of the parse tree of an Arabic sentence after applying the steps one and two.

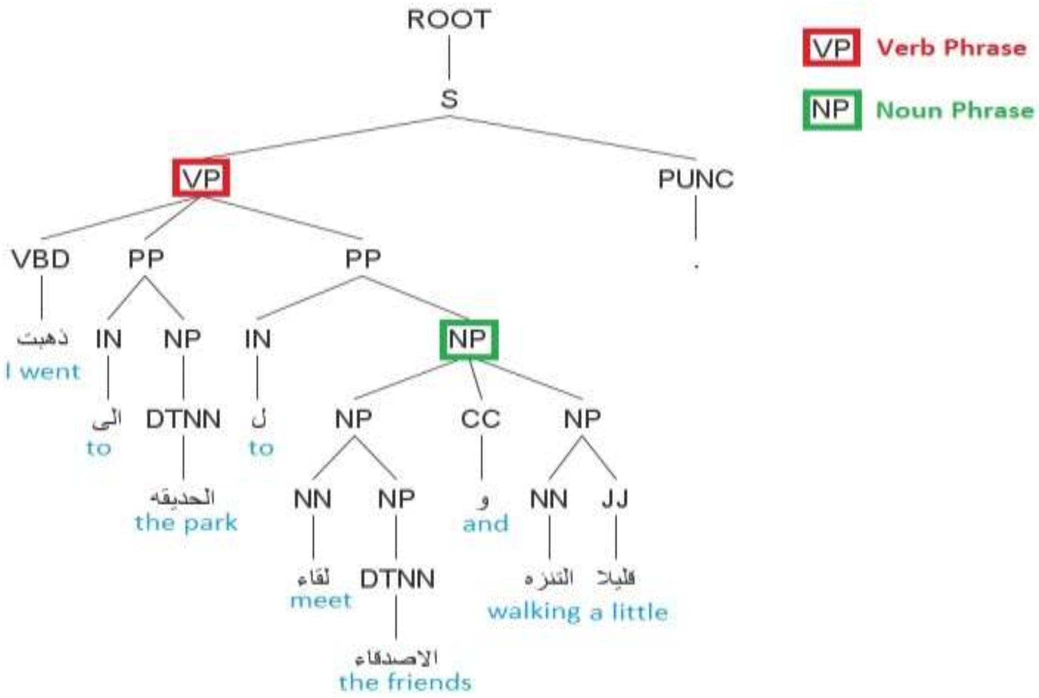


Figure 3.8 An example of Arabic sentence parse tree with PoS tags which generated by using both of Stanford Arabic Parser and Tagger tools

Step 3. For the noun clauses, the algorithm looks at the word sequence of each (Bigram) word pair and adds the words matching the following rule (pattern) to the candidate list of concepts.

- [Noun+Noun] → add [Noun+Noun] pattern to the list of candidate concepts.
- [Noun+Adjective] → add [Noun+Adjective] pattern to the list of candidate concepts.
- [Adjective+Noun]→ add [Adjective+Noun] pattern to the list of candidate concepts.
- [Noun+Stopword] → add only the [Noun] to the list of candidate concepts.
- [Stopword + Adjective] → add only [Adjective] to the list of candidate concepts.

Step 4. for each one of the verb clauses the algorithm adds words that match the [Verb+Object] rule to the candidate list of concepts. In the first step, in order to determine the Object that related to the Verb in the sentence, the Stanford dependency parser "dependency analyzer" is used [114]. Stanford dependency parser is used to specify and identify grammatical dependency relationships among the words in the phrase. The Arabic language dependency dataset "universaldependencies.org" is used to train the Stanford dependency parser in order to use it for the Arabic dependency analysis.

After analyzing the sentence with the “Stanford dependency parser”, the words that match the [Verb+Object] rules are added to the candidate list of concepts.

In some cases, the “Stanford dependency parser” may not be able to identify the object because of the limitations in the Arabic dependency dataset. If this is the case, the grammar structure of the standard Arabic sentence ([Verb + Subject] + Object) is used to find the object in the sentence [44]. When the Object is identified then the Object added to the candidate concepts list with the Verb related to it. The generated final list of the candidate concepts is called Bag of Concepts.

Our SP concept extraction algorithm is tested against randomly selected 100 Arabic sentences with manually extracted concepts. For these 100 Arabic sentences, the concepts that extracted by the proposed concept extraction algorithm are compared with the manually extracted concepts. The comparison results show that the concept extraction algorithm is achieved an accuracy value of 97% for concept extraction compared to the manual method. Table 3.2 shows an example of Arabic sentences and the concepts extracted from these sentences using our concept extraction algorithm.

Table 3.2 An example of concepts that extracted from Arabic of sentences using our SP concept extraction algorithm

1	**Arabic sentence**	إن أسعار الذهب بالغة الحساسية لتحركات أسعار الفائدة العالمية
	English translation	Gold prices are very sensitive to movements of the global interest rate.
	Grammatical parse	(ROOT (S (VP (VBP إن) (NP (NN أسعار) (NP (DTNN الذهب))) (ADJP (JJ بالغة) (NP (DTNN الحساسية))) (PP (IN ل) (NP (NNS تحركات) (NP (NN أسعار) (NP (DTNN الفائدة) (DTJJ العالمية))))))))
	Extracted concepts	سعر ذهب / الحساسية لتحركات / تحركات أسعار/ الفائدة العالمية / حساسة لحركة سعر الفائدة
2	**Arabic sentence**	يمكن أن يكون المحيط شيآ في غاية التعقيد.
	English translation	The ocean can be a very complicated thing.
	Grammatical parse	(ROOT (S (VP (VBP يمكن) (NP (DTNN أن)) (S (VP (VBP يكون) (NP (DTNN المحيط) (DTJJ أمرا)) (PP (IN في) (NP (NN غاية) (NP (DTNN التعقيد)))))))))
	Extracted concepts	المحيط / المحيط شيء / شيء معقد / شيء / غاية التعقيد / المحيط معقد
3	**Arabic sentence**	تغمرنا السعادة بهذا التكريم باختيارنا أفضل مدونة صحفية بالعربية.
	English translation	We 're thrilled to be honored as the jury 's choice for the Best Journalistic Blog in Arabic.
	Grammatical parse	(ROOT (S (VP (VBP تغمر) (NP (PRP نا)) (NP (DTNN السعادة)) (PP (IN ب) (NP (NP (DT هذا)) (NP (DTNN التكريم)))) (PP (IN ب) (S (VP (VBG اختيار) (NP (PRP$ نا)) (NP (NN أفضل) (JJ مدونة) (JJ صحفية)) (PP (IN ب) (NP (DTNN العربية)))))))))
	Extracted concepts	السعادة باختيارنا / أفضل مدونة / مدونة صحفية / مدونة صحفية بالعربية / العربية

3.4.2. Feature Extraction and Representations

Since ML approaches are considered as the main core of our proposed Sentence-level concept-based Arabic SA approach, which are responsible for identifying and deciding the sentiment polarity of the input sentence so that the input sentence must be transformed into a set of numerical features that can be useful for the ML algorithm. In this work, we present and exploit a different feature extraction and representation techniques, to extract a variety of features sets from the input sentence and then fed them as input to ML decision model. These feature set are concept-based features, lexicon based features, Bag of Word features and Word2Vector features.

1- The Concept-based Features (CBF) Includes;

- SenticNet features (The number of concepts extracted from the sentence and found in our generated Arabic SenticNet, The summation of the extracted concepts scores which are obtained from Arabic SenticNet).

- Part of speech (PoS) features (The number of nouns, adjectives, and adverbs found in the sentence).

- Modification features (This binary feature is set to 1 if the sentence has any word modified by an adverb, adjective, or noun otherwise it’s set to 0).

- Negation features (The negation binary feature determines whether there is any negation in the sentence).

2- The Lexicon Based Features (LEX) Includes;

- Lexicon features (Positive words number, Negative words number, Positive words number divided by the negative word number, the sum of the positive scores and the sum of the negative scores). The version of Arabic SentiWordNet that called ArSneL [111] is used to extract these features from the sentence.

3- The Bag-of-Word Features (BoW) Includes;

- Bag-of-Word features (the sentence is represented using either Uni-grams or Bigrams features as a vector, and these features are weighed using TF-IDF method).

4- The Word2Vector Features (W2V) Includes;

- Word2Vector features (each word within the sentence is transformed into a real-valued 300-dimensional vector, the Word2Vector features are generated by the aggregation of the vectors of each word in the sentence). Word Vectors can be obtained from Word2Vec model. We used our large-scale corpus GLASC to train and generate the Word2Vec model[1].

Similar to the approach presented in Section 3.2, ML algorithms can be also utilized for the concept-based SA task. In such case, the SA task of a sentence can be considered as classification and regression problems. In this work, both ML classification and regression approaches are considered to identify the sentiment category and predict the sentiment score of the input sentence respectively.

For the sentiment classification task, we employ four different ML classifiers such as (SVM [15], HMM [69], NB [107], and LR [71]) to generates the sentiment classification model. Moreover, different versions of combined classifiers methods such as (SVM-LR, SVM-NB, and SVM-HMM) which are ensembled using stacking technique, are also taken into our considerations. For the sentiment regression task, we use three types of ML regression algorithms such as (SVR [110], MLR [111] and MLP [112]) to build SA regression model. These classification and regression algorithms can be trained using a sentence based dataset that is can be generated from our Arabic large-scale SA corpus GLASC. The processes of generation this dataset using a different combination of features set will be presented in the next chapter together with our evaluation process for the proposed sentiment classification and regression models.

[1] "https://code.google.com/p/word2vec/"

4. DATA EVALUATION AND TESTING

In this chapter, we presented different sets of evaluation and validation experiments to our proposed and used approaches in this thesis. In Section 4.1 we presented two different approaches to evaluate the quality of our generated large-scale GLASC corpus, based on statistical measures and Zipf law distribution. In Section 4.2 we introduced four different benchmark datasets, which are generated from our GLASC corpus and, used in the evaluation experiments for our proposed ML-based SA approaches for Arabic language based on the evaluation metrics discussed in Section 4.3. in Section 4.4 we carried out a comprehensive experiments for evaluating our proposed Arabic document-level SA system and then we provided a comparative discussion of the obtained results. in Section 4.4. a similar type of the extensive evaluation experiments are also applied for our proposed concept-based Sentence-level Arabic SA system and then we provided a comparison and discussion of the obtained results. In this section, we also measured the quality of our Ar-SenticNet concept-based sentiment lexicon for Arabic, which is generated by translating the English version of SenicNet to Arabic using our proposed translation and extension approach. The quality evaluation of Ar-SenticNet is done by measuring its concept coverage over our large-scale corpus GLASC.

4.1. GLASC Corpus Quality Evaluation

In order to evaluate the quality of our GLASC, we obtained two types of corpus quality evaluation tests. The first test uses statistical measures to evaluate the corpus quality and the second test uses Zipf law to measures the corpus quality based on Zipf distribution.

4.1.1. Corpus Statistics Evaluation

We obtained different statistical measurements related to the produced Arabic Corpus such as the number of files in each category, total number of words, the average number of words in each file, total number of unique terms, total number of unique terms in the corpus, the average number of unique terms in each file, total number of sentences and the average number of sentences in each file as shown in Table 4.1.

Table 4.1 Statistics of our large-scale Arabic corpus

Category	Negative	Positive	Neutral
Number of files	266,376	225,628	218,310
Total number of words	91,051,658	70,596,129	51,061,595
The average number of words in each file	342	313	234
Total number of unique terms	155,929	154,336	156,752
Total number of unique terms in the corpus		230,123	
The average number of unique terms in each file	204	184	142
Total number of sentences	4,567,333	3,550,913	2,575,378
The average number of sentences in each file	17	16	12

Since the Tone value provided by GDELT is used to assign the news articles files into three various categories (positive, negative and neutral) to obtain our corpus, we need to evaluate the quality of this file assignment. For this task, we considered using ArSenL [59] which is a large-scale Standard Arabic sentiment and opinion-mining lexicon contains a total of 28,760 Arabic lemmas with corresponding sentiment scores.

In the first test, we calculated the average number of the positive, negative and neutral terms in each category of our GLASC corpus using ArSenL as shown in Figure 4.1.

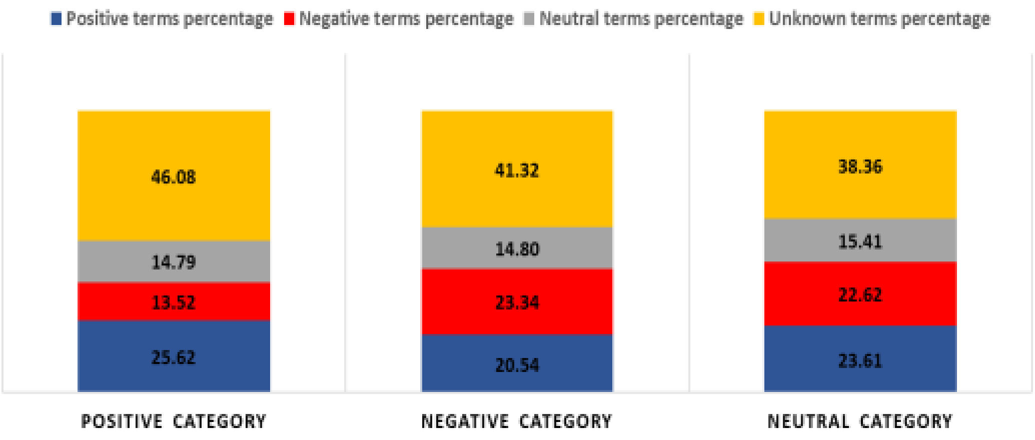


Figure 4.1 The average number of the positive, negative and neutral terms in each category

The results shown in Figure 4.1, verified that the positives, negative and neutral terms ratio in a specific category are compatible with the nature of that category. These results can be summarized as following; (i) The positive category of our GLASC corpus contains a (25.62%) percentage of the positive terms which is greater than the percentage of both negative and natural terms. (ii) In the negative category of our corpus, the percentage of negative terms is (23.34%) which is greater than the percentage of the positive and the natural terms in this category. (iii) For the neutral category of our corpus, the percentage of the negative terms is (22.62%) and the positive terms is (23.61%), which are closely similar to each other, however, the percentage of the negative and positive terms is greater than the percentage of the neutral terms in this category.

This result can conform the quality of our corpus where the positive, negative and neutral terms are harmoniously distributed in each category.

The second test we performed over the corpus is done by calculating the average positive and negative scores for the terms in each category of our corpus using ArSenL as shown in Figure 4.2.

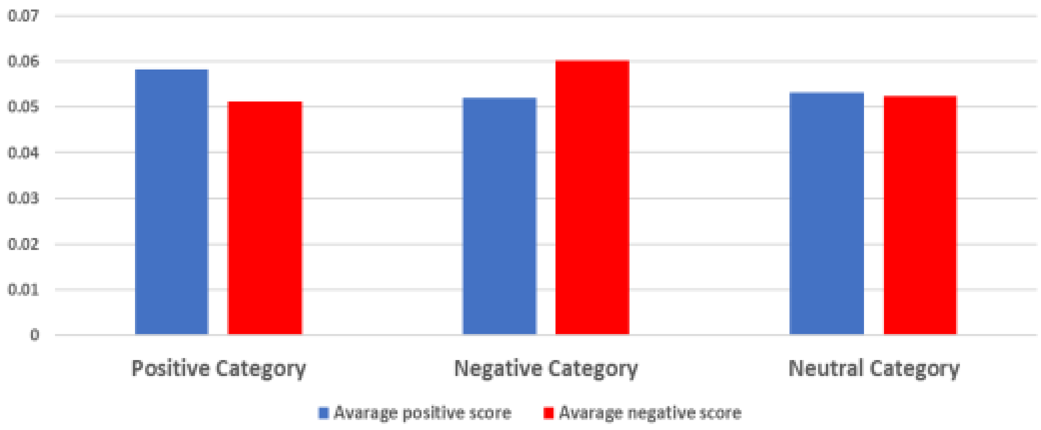


Figure 4.2 The average score values of the positive and negative terms that found in each category

The results in Figure 4.2 shows that; (i) The average of positive terms scores is greater than average of negative terms scores in the positive category. (ii) In the negative category, the average of negative terms scores is greater than average of negative terms scores. (iii) In the neutral category, the average scores of both positive and negative terms are very close to each other.

These results are also confirming the quality of our GLASC corpus where the ratio of positive and negative terms scores is compatible with the category that contain these terms.

4.1.2. Zipf's Law

The Zipf's law has been formulated as an empirical law which is revealed using mathematical statistical knowledge. It was named after being published by George Kingsley Zipf, professor of linguistics at Harvard University in the United States in 1930. This empirical law is about the frequency of words found in a text written in any human language. In 1949, the linguist George Zipf was aware of something strange about the frequency of use of certain words. According to the findings of Zipf, the vast majority of words were seldom used, but a few words were always used. A striking pattern emerged when words were ordered by frequency of use. The word in the first order was always used twice as frequently as the second word, and the third word was used as often as three times. He found that this rule named as an order-frequency rule could be used to express income distributions in any country so that the wealth of the richest person was twice as much as the next rich one, and so on.

Zipf's first law: When observations frequencies of words within a document are classified from minor to major, the observance frequencies *(f)* and order number *(r)* and the numerical values obtained by multiplying *(c)* are approximately described as constant (Eq 4.1).

$$f \propto \frac{1}{r} \; and \; f.r = c \qquad (4.1)$$

Mandelbrot (Manning and Schütze, 2003) [26] showed that the generalization given by Zipf is actually very bad at the point of detailing when we are working with larger arrays. Mandelbrot changed the general relation between order and frequency as in Equation 4.2, which would be more appropriate for the experimental distribution of words.

$$\log(f) = \log(p) - B \log(r + \rho) \qquad (4.2)$$

In Equation 4.2. *P, B* and ρ are the parameters of the text, and they always reveal the richness of the vocabulary used in the text together. The original associative hyperbolic distribution given by Zipf (Eq 4.1) applies to (Eq 4.2) as well. When the statement is given in Equation 4.2. is transformed using logarithm scale axis line, the order *(r)* of the slope for maximum value conforms to a line with a slope of -*B*. If it's *B* = 1 and ρ = 0 in the equation, it's seen that it will be equal to the statement given in Equation (Eq 4.1) for Zipf's first law.

Zipf law can be used to measure the corpus quality by measuring its words frequency distribution correctness [115]. In order to verify our corpus quality, we calculated the word frequency distribution of our corpus and then calculating it's fitting to Zipf's law distribution as shown in Figure 4.3.

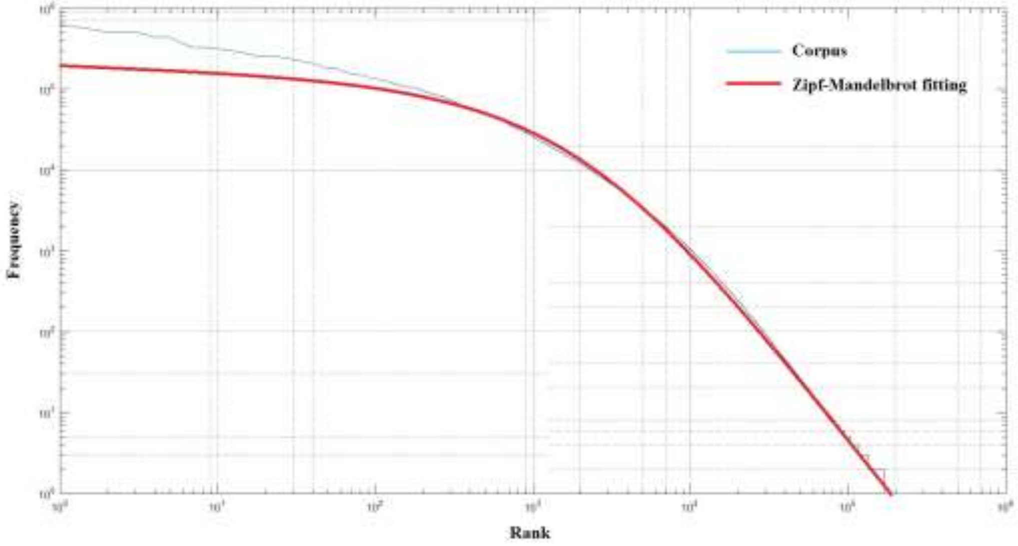


Figure 4.3 The rank-ordered frequency distribution of our corpus and its fitting with regard to the Zipf-Mandelbrot law

From the result in the Figure 4.3, it can be shown that the data of our GLASC can be fit well by Zipf-Mandelbrot law distribution with coefficient values of P=22, B=1.15, $\rho = 25$.

4.2. Dataset Generation for Evaluating the Machine Learning Based Sentiment Analysis Systems

large-scale Arabic sentiment corpus is used for generating different datasets that will be used for training the ML classifier for SA. Figure 4.4 shows the process of generating the datasets which are consist of the following steps:

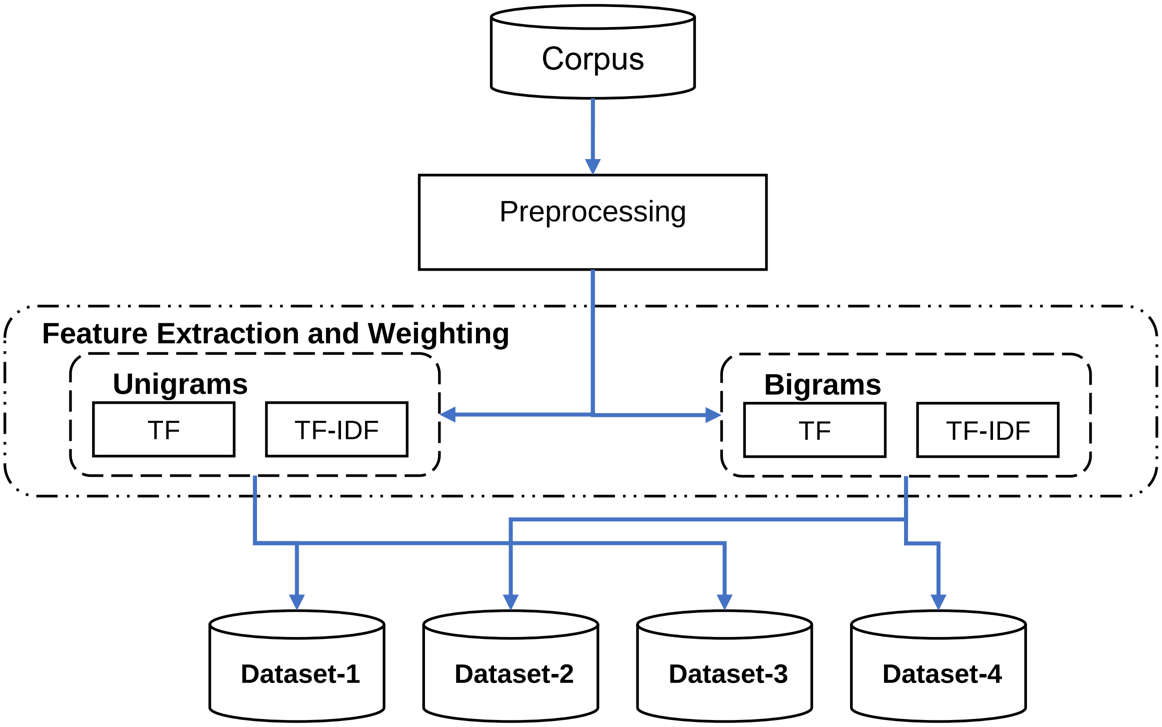


Figure 4.4 Datasets generation process

4.2.1. Prepressing

The first stage is the preprocessing where each document in GLASC is processed by tokenizing all the terms and apply normalization and removing the stop words, then finding the root for each token using Buckwalter morphological analyzer's Ara-morph Arabic lemmatizer [106].

4.2.2. Feature Extraction and Weighting

The second stage is the extraction and weighting of features. Two different feature extraction methods based on unigrams and bigrams are used. After calculating all the terms vectors for each document in the corpus, the dataset can be represented as a matrix where documents represent the rows to and words feature (in our case TF and TF-IFD) represents the columns.

For the system evaluation, four different datasets are generated using two different feature extractions (unigrams and bigrams) and two different feature weighting methods (TF and TF-IDF). Table 4.2 provides a summary of each generated dataset properties.

Table 4.2 The properties of the generated datasets

	Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
Features type	Unigrams	Unigrams	Bigrams	Bigrams
Features weighting	TF	TF-IDF	TF	TF-IDF
Number of features	230,123	230,123	5,600,000	5,600,000
Train instances	***Negative***	***Positive***	***Neutral***	
	186,463	157,940	152,817	
Test instances	***Negative***	***Positive***	***Neutral***	
	79,913	67,688	4,5845	

4.3. System Evaluation Metrics

In order to evaluate the performance of our proposed ML-based sentiment classification and regression models, we considered using 10-fold cross-validation method to calculate the classification accuracy and F-score for the ML sentiment classification models, and the prediction MAE and RMSE for the ML sentiment regression models. These evaluation measures are explained as following [116] [72];

4.3.1. Sentiment Classification Model Performance Evaluation Metrics

In this work, we considered two classification model performance evaluation metrics that are most commonly used to evaluate the performance of the ML-based classification models, which are the classification accuracy and the F-score measures. These classification model performance evaluation metrics can be calculated as following [116] [117];

4.3.1.1. Confusion Matrix

This matrix containing the predicted and actual class values of the document. The representation of this matrix is given in Table 4.3.

Table 4.3 The Confusion Matrix

	P’ (Classified)	**N’ (Classified)**
P (Actual)	TP	FN
N (Actual)	FP	TN

4.3.1.2. Accuracy

It’s the measure of how the accurate is a classifier performs the correct class assignments. In other words, it’s the value expressing at what proportion the assignments performed by classifier are accurate. The classification accuracy measure answers the question of "How correct is the classifier in classifying all samples?" and it’s calculated as:

$$\text{Accuracy} = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{4.3}$$

Where *c* refers to the number of classes and True Positive (TP) and True Negative (TN) is number of correctly classified instances by the model, False Positive (FP) and False Negative (FN) refers to the number of instances that miss-classified by the model.

4.3.1.3. Recall

The recall is the measure of how much of the assignments that the classifier makes to the appropriate class. This measure could be used as an answer to the question of “How much of the samples in a class are classified correctly?” and it’s calculated as;

$$\text{Recall} = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TP_i + FN_i} \tag{4.4}$$

4.3.1.4. Precision

It’s the measure that answers the question of "How sensitive is a classifier in the classification that made for a class?" In other words, "at what proportion the accurate result is obtained for that assignment class?". The precision of a classification model is calculated as;

$$\text{Precision} = \frac{1}{c}\sum_{i=1}^{c}\frac{TP_i}{TP_i + FP_i} \tag{4.5}$$

4.3.1.5. F-score

It's a measure of the classification accuracy that takes into account both the recall and the precision. F-Score that will demonstrate the harmonic mean of Precision and Recall of a classification model and it's calculated as;

$$\text{F} - \text{score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4.6}$$

4.3.2. Sentiment Regression Model Performance Evaluation Metrics

For our sentiment regression model performance evaluation task, we considered two evaluation metrics that commonly used to evaluate the ML regression models. These metrics are the Mean Absolute Error (MAE) and the Root Mean-Squared Error (RMSE) [117];

4.3.2.1. Mean Absolute Error (MAE)

It's the average of the difference between the estimated value by the regression model and the actual value in all test cases. The formula required for the calculation of MAE is as shown in Equation (Eq 4.7):

Let's assume that the actual value is *a*, and the estimated value is *c*

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - c_i| \tag{4.7}$$

4.3.2.2. Root Mean-Squared Error (RMSE)

RMSE is often used as a calcualte of the difference between predicted values by the estimator model and actual values. RMSE is simply the square root of the MSE which is calculated by taking the average of the squares of the difference between each predicted value and its corresponding actual value.

RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - c_i)^2} \qquad (4.8)$$

Assuming that the actual value is *a*, and the estimated value is *c*

MAE and RMSE are computed for each ML regression algorithm. MAE and RMSE are used together to describe the errors variation in a set of estimates. MAE and RMSE are negative-oriented so that the lower values of MAE and RMSE represents better performance. Therefore, the algorithm with low MAE value is considered as the best algorithm. RMSE value must be greater than or equal to the value of MAE.

4.4. Machine Learning Approach for Document-level Arabic Sentiment Analysis Evaluation

In this section, we carried out comprehensive and comparative experiments for evaluating our proposed Arabic document-level SA system which is based on the ML classification and regression approaches. Our goal with this evaluation experiments is to determine and specify the best performing ML classification and regression models to use them for the proposed Arabic document-level SA system.

4.4.1. Sentiment Classification Model Generation and Testing

4.4.1.1. Experiments with Base Learners

In the first experiment, we used the four sentiment datasets generated in Section 4.2 earlier individually, for training the five base learners (SVM, HMM, NB, ANN, and KNN).

The 10-fold cross validation method is considered to reduce the influence of variability in the training dataset. So that each individual classification model is trained on 70% of randomly drawn samples from the original dataset, and tested on remaining 30% samples repeatedly for ten times, and each time when testing dataset is applied the classification accuracy and F-score performance metrics are calculated. At the end of the 10 folds, the average value of the calculated accuracy and F-score performance metrics are obtained. This process is applied to each individual base learner using each one of the four datasets separately. The obtained values of the accuracy and F-score performance metrics for each base learner using four different datasets is shown in Table 4.4.

Table 4.4 The classification Accuracy (A) and F-score (F) for base learners using four different datasets

		Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
SVM	A	0.8525	0.8547	0.8699	**0.8776**
	F	0.8491	0.8527	0.8662	**0.8706**
HMM	A	0.8271	0.8457	0.8692	0.8675
	F	0.8183	0.8352	0.8629	0.8662
NB	A	0.7730	0.7755	0.7838	0.8008
	F	0.7726	0.7682	0.7832	0.7905
ANN	A	0.7212	0.7229	0.7375	0.7407
	F	0.6725	0.7226	0.7284	0.7396
KNN	A	0.5503	0.5655	0.6501	0.6671
	F	0.4725	0.5128	0.5716	0.6109

According to our experiments, for the all the four datasets the best classification performance is achieved by both of SVM and HMM classification methods. SVM classifier provided its best performance of 87.76% F-score by using the dataset with Bigram features and TF-IDF weights followed by HMM classifier provided its best performance of 86.75% F-score using also the same dataset.

To identify the best base leaner classifier method among (SVM, HMM, NB, ANN, and KNN) we calculate the average value for the classification accuracy and F-score for each (SVM, HMM, NB, ANN, and KNN) classification method over all our four datasets as shown in Figure 4.5.

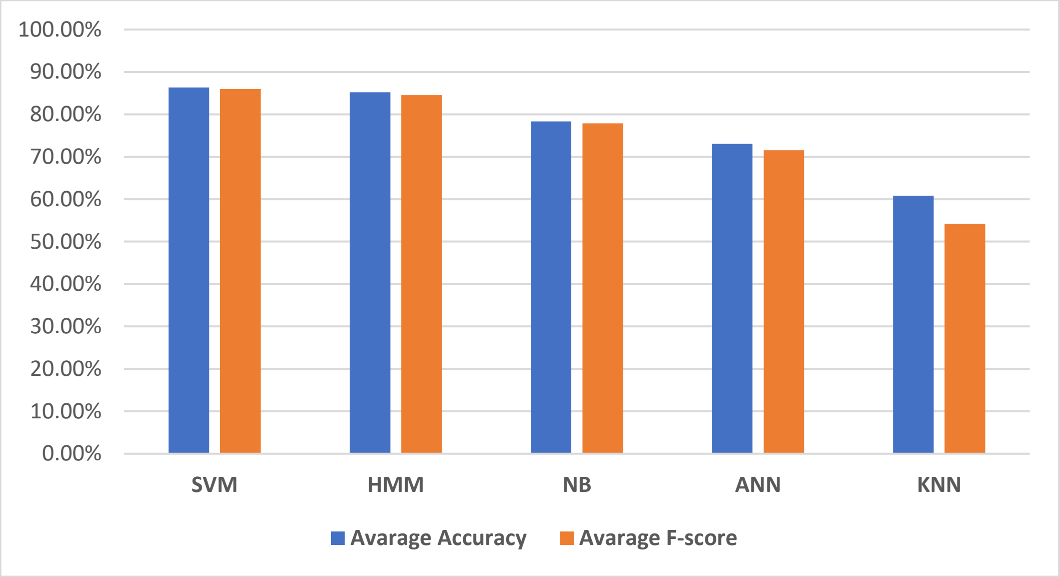


Figure 4.5 The average value of accuracy and F-score of each base learner for all 4 datasets

The results in Figure 4.5 clearly shows that the SVM base learner classification method has the best classification performance of 85.96% of average F-score, over the other base learners followed by HMM base learner which achieved a classification performance of 84.57% of average F-score.

In addition to the base classifiers evaluation in the previous experiment, we consider another set of experiments to evaluate ensemble classifiers with the same four datasets and evaluation metrics.

4.4.1.2. Experiments with Classifier Model Ensemble

We used three different classifier model ensembling methods which are Bagging, Boosting, Random Subspace and stacking. These methods used to combine the same type of classifiers models for each one of (SVM, HMM, NB, ANN, and KNN) base learners in order to increase the classification performance.

Bagging method generates 5 different bootstrap training subsets which are drawn from the original training set with replacement. These five training subsets are used to train five model form similar type of classifier for each of (SVM, HMM, NB, ANN, and KNN). The final prediction is generated by the majority voting of these five models.

Table 4.5 shows the accuracy and F-score results for each ensemble classifier using bagging method after training and evaluating with our four sentiment datasets using 10 fold cross-validation.

Table 4.5 The classification Accuracy (A) and F-score (F) for ensemble learners using Bagging method for four different datasets

		Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
SVM	A	0.8595	0.8578	0.8724	**0.8839**
	F	0.8574	0.8548	0.8688	**0.8793**
HMM	A	0.8332	0.8418	0.8731	0.8740
	F	0.8254	0.8290	0.8679	0.8733
NB	A	0.7634	0.7753	0.7863	0.8137
	F	0.7629	0.7668	0.7851	0.8072
ANN	A	0.7253	0.7340/	0.7437	0.7486
	F	0.6804	0.7335	0.7330	0.7476
KNN	A	0.5569	0.5562	0.6621	0.6701
	F	0.5037	0.5124	0.5936	0.6175

Boosting used to train five classification model form similar type of classifier for each of (SVM, HMM, NB, ANN, and KNN) in sequence. Each classification model is focused on the misclassified samples by the preceding model. Similar to bagging the final classification decision made by majority voting of these five models. The accuracy and

F-score results of each ensemble classifier based on boosting method after training and evaluation using our four sentiment datasets using 10 fold cross-validation, are shown in Table 4.6.

Table 4.6 The classification Accuracy (A) and F-score (F) for ensemble learners using Boosting method for four different datasets

		Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
SVM	A	0.8578	0.8330	0.8876	**0.8837**
	F	0.8548	0.8285	0.8821	**0.8765**
HMM	A	0.8418	0.8194	0.8771	0.8799
	F	0.8290	0.8138	0.8698	0.8793
NB	A	0.7753	0.7754	0.7952	0.8013
	F	0.7668	0.7745	0.7944	0.7930
ANN	A	0.7340	0.7305	0.7398	0.7443
	F	0.7335	0.6820	0.7338	0.7436
KNN	A	0.5562	0.5551	0.6532	0.6730
	F	0.5124	0.5047	0.5854	0.6133

Random subspace method is similar to bagging method in concept, however, its trained five similar models for each classification method on the same dataset with random feature subspaces, where each random subspace contains 50% of the available feature space. Table 4.7 shows the accuracy and F-score results for each ensemble classifier using random subspace method after training and evaluating with our four sentiment datasets with 10 fold cross-validation.

Table 4.7 The classification Accuracy (A) and F-score (F) for ensemble learners using random subspace method for four different datasets

		Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
SVM	A	0.8605	0.8692	0.8907	**0.9057**
	F	0.8578	0.8675	0.8858	**0.9021**
HMM	A	0.8464	0.8515	0.8898	0.8926
	F	0.8354	0.8391	0.8839	0.8921
NB	A	0.7852	0.7868	0.7788	0.8070
	F	0.7843	0.7821	0.7783	0.7992
ANN	A	0.7185	0.7404	0.7459	0.7520
	F	0.6775	0.7399	0.7395	0.7511
KNN	A	0.5674	0.5814	0.6635	0.6884
	F	0.5185	0.5310	0.5992	0.6376

The experimental results for (Bagging, Boosting and Random subspace) ensemble methods, shown in Tables 4.5–4.7 can be summarized as following: the classification accuracy results achieved by Bagging, Boosting and Random subspace ensemble methods are higher than the classification accuracy results achieved by the base learners. For all of the ensemble methods that used, the best classification accuracy is achieved by SVM and HMM classifiers using the dataset with Bigram features and TF-IDF weights (Dataset-4). Random subspace ensemble method has achieved the highest classification accuracy over the other Bagging and Boosting ensemble methods. The best explanation for this phenomenon is that most of the learning algorithms are sensitive to the dimensionality of the training data in a negative manner and, since sentiment classification problem has a high dimensional feature space data that may contain noisy features which may lead to overfitting problem. Since Random subspace ensemble is based on feature partitioning, so it can reduce the risk of overfitting problem and improve the classification performance.

In order to identify which classifier ensemble method has achieved the best classification performance for each of (SVM, HMM, NB, ANN, and KNN) methods, we calculate the average value of the classification accuracy and F-score for each classification method based on using three types of the ensemble (Bagging, Boosting and Random subspace) method, for all of our four datasets as shown in Figure 4.6.

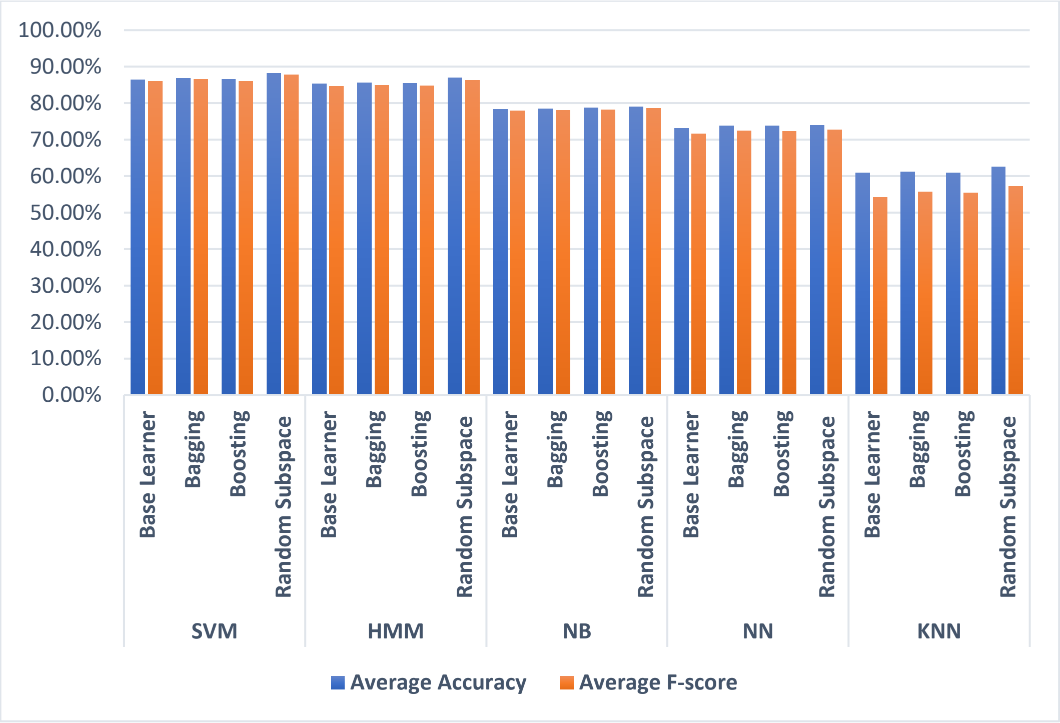


Figure 4.6 The average value of accuracy and F-score of each base learner and ensemble learner using Bagging, Boosting and Random subspace, for all 4 datasets

For the results shown in Figure 4.6, it can be shown that for all (SVM, HMM, NB, ANN and KNN) classifiers the bagging and boosting ensemble method provided a small value of improvement in classification performance in terms of the accuracy and F-score, compared to the value of improvement in classification performance that achieved by Random subspace ensemble method.

To show the impact of ensemble method on the classification performance, we also calculated the percentage value of the average improvement in classification performance for (SVM, HMM, NB, ANN, and KNN) classifiers using Bagging, Boosting and Random subspace ensemble methods for all our datasets as shown in Figure 4.7.

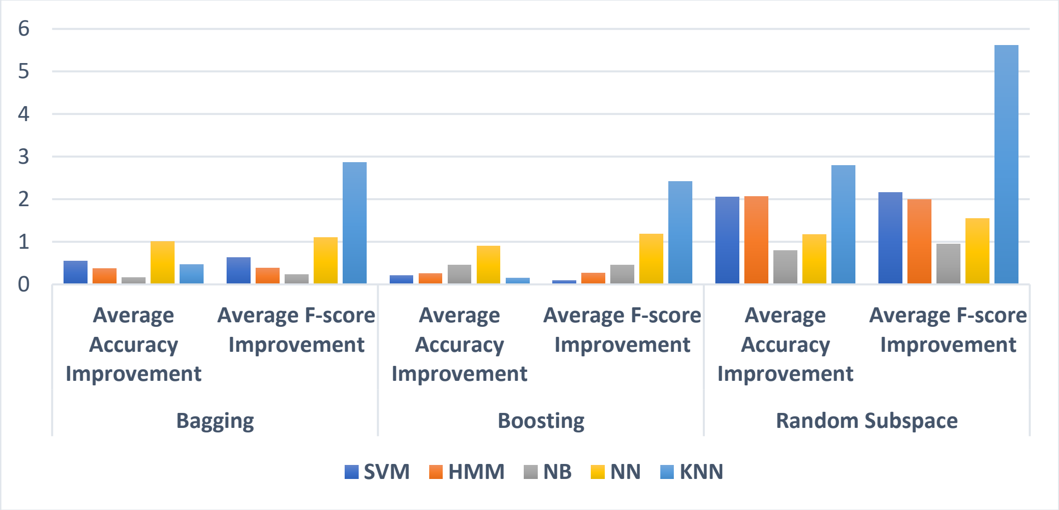


Figure 4.7 The percentage of improvement in the average of classification accuracy and F-score for each base learner using Bagging, Boosting and Random subspace ensemble techniques for all 4 datasets

The results shown in Figure 4.7 shows that the Random subspace ensemble method has more impact in improving the classification performance for all (SVM, HMM, NB, ANN, and KNN) classifiers over the other Bagging, Boosting methods. The results also show that these Bagging, Boosting and Random subspace ensemble methods have an impact on improving the classification performance of the weaker classifiers which is in our case KNN, where its base classification F-score improved by 2.8%, 2.4% and 5.6% using Bagging, Boosting and Random subspace ensemble methods respectively.

4.4.1.3. Experiments with Classifier Model Fusion

Stacking is a classification model fusion method which is concerned with combining multiple classifiers generated by using different learning on a single dataset. This method implies two stages, the first stage consists of training different classification models called base-level classifiers. In the second stage, a meta-level classifier is learned that combines the outputs of the base-level classifiers and the predictions of base learners (level-0 models) are used as input for meta-learner (level-1 model).

We performed different batches of experiments, wherein each experiment we used a different combination of classifiers methods such as (SVM+NB, NB+HMM, NN+SVM and, SVM+HMM) as level-0 models. For the level-1 model, we used a multilayer perceptron MLP as meta-classifier to combine the decisions of the level-0 classifier models. Table 4.8 shows the accuracy and F-score results for each classifier combination using the stacking method after applying 10-fold cross validation for training and evaluating with our four sentiment datasets.

Table 4.8 The classification Accuracy (A) and F-score (F) for combined learners using stacking method for four different datasets

		Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
SVM+NB	A	0.8633	0.8599	0.8811	0.9144
	F	0.8525	0.8599	0.8775	0.9142
NB+HMM	A	0.8263	0.8515	0.8794	0.8851
	F	0.8184	0.8392	0.8779	0.8851
ANN+SVM	A	0.8589	0.8651	0.8750	0.9071
	F	0.8517	0.8582	0.8733	0.9057
SVM+HMM	A	0.8662	0.8863	0.8974	**0.9238**
	F	0.8661	0.8839	0.8967	**0.9235**

The results in Table 4.8 can show that by using stacking method we were able to improve the accuracy of the all combined (fused) classification models rather than using the induvial models. The highest classification accuracy of 92.35% F-score is achieved by SVM+HMM classifiers fusion, followed by 91.42% F-score by SVM+NB classifiers fusion, using the dataset with Bigram features and TF-IDF weights.

In order to identify which combination of the fused classification method can achieve the best classification performance, we calculate the average value of the classification accuracy and F-score for each classifier combinations, over the all of our four datasets as shown in Figure 4.8.

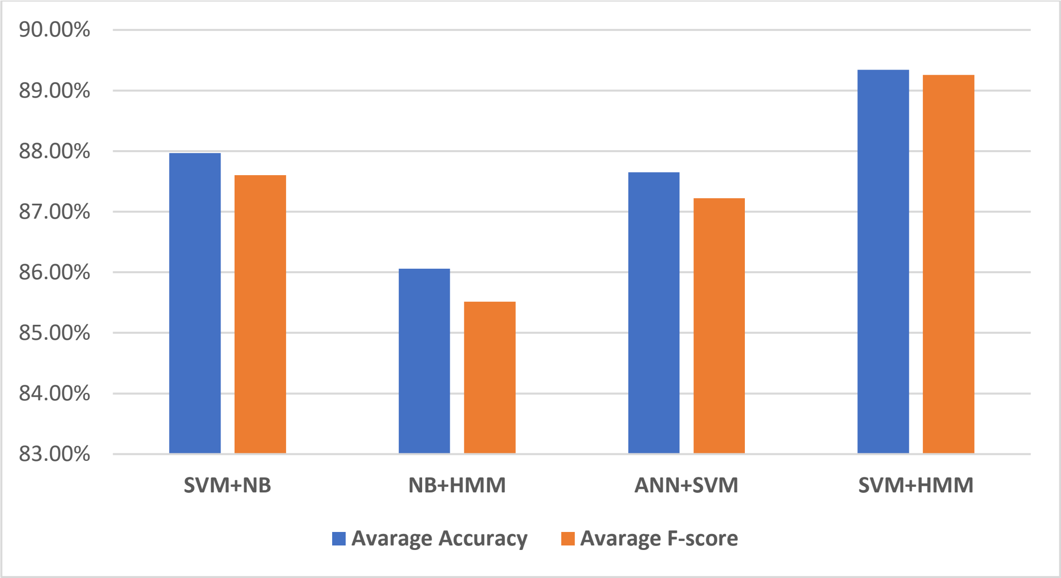


Figure 4.8 The average value of accuracy and F-score of each combined learner using stacking method for all 4 datasets

The results in Figure 4.8 clearly shows that best classification performance can be achieved by the (SVM-HMM) classifiers combination followed by (SVM-NB, ANN-SVM, and NB-HMM) classifiers combination.

In order to show the impact of using classifier combination in the regard of improving classification performance over the using individual classifiers, we calculate the percentage value of the improvement in the accuracy and F-score using the combined classifiers instead of using each classifier individually, and results are shown in Figure 4.9.

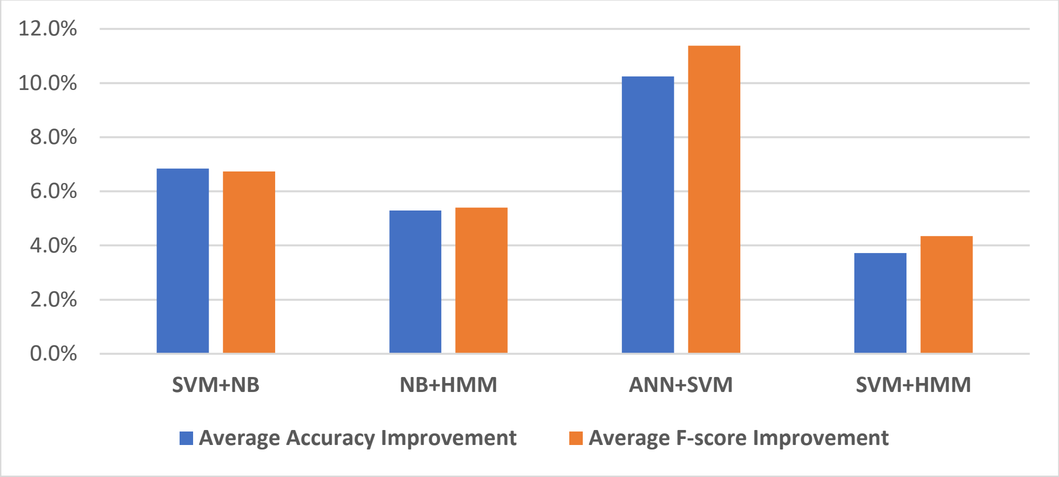


Figure 4.9 The percentage of improvement in the average of classification accuracy and F-score for each combined learner using stacking method for all 4 datasets

Although the best classification performance is achieved by using (SVM-HMM) classifiers combination, the result in Figure 4.9 shows that using (ANN-SVM) classifiers combination is provide 11.3% improvement in F-score over the average performance achieved using (ANN or SVM) classifier individually, which mean that the stronger SVM classifier is able to enhance the performance of the weaker ANN classifier when it used together. The only 4.3% improvement in F-score is achieved using the combination of the two stronger classifiers SVM and HMM (based on our results) this can be explained by, both of SVM and HMM models are classifying most of the testing samples similarly. In other words, SVM and HMM classifier models make the same decisions for most of the testing samples. So that combining these two classifier models results in a small improvement in the classification performance.

Since we used 4 different datasets with different features in order to evaluate our sentiment classification system, we investigated the impact on the classification performance from the perspective of using different feature sets. We first selected the Dataset-4 (with Bigram features and TF-IDF weights) as a baseline for our experiment, then we calculated the percentage of the improvement in F-score for each classifier/classifier-combinations when the baseline dataset is used instead of the other datasets. the percentage of improvement in F-score can be calculated as;

$$\text{Improvement in Fscore} = \frac{\text{Fscore}_{baseline} - \text{Fscore}_{compared}}{\text{Fscore}_{compared}} \quad (4.9)$$

The results in Figure 4.10 shows the percentage of improvement in F-score for each classifier, when Dataset-4 is used instead of the Dataset-3 (with Unigram features and TF weights), Figure 4.11 shows the percentage of improvement in F-score, when Dataset-4 is used instead of the Dataset-2 (with Bigram features and TF-IDF weights) and Figure 4.12 shows the percentage of improvement in F-score, when Dataset-4 is used instead of the Dataset-1 (with Bigram features and TF weights).

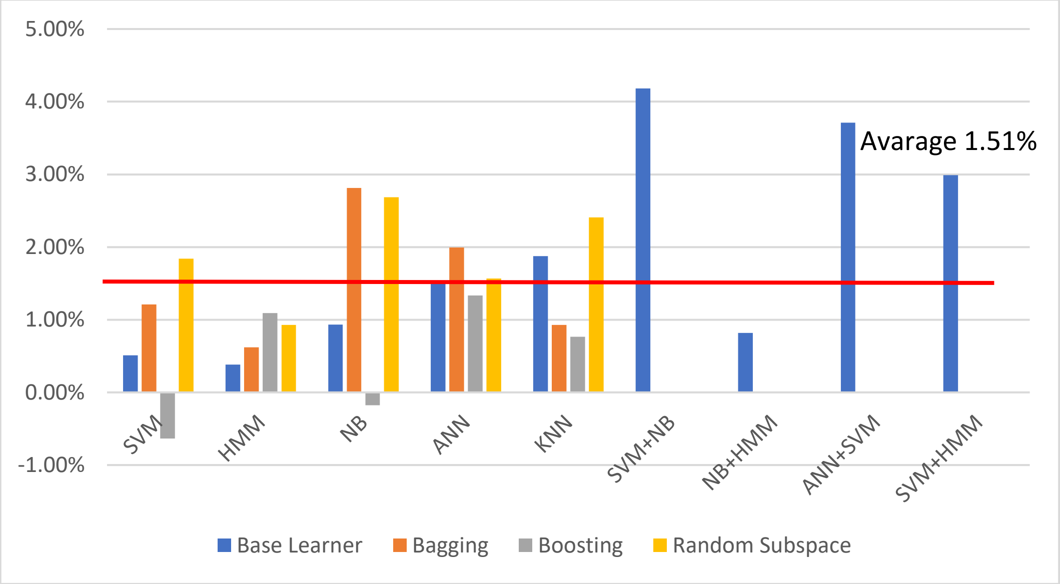


Figure 4.10 The percentage of improvement in classification accuracy F-score that achieved using Dataset-4 instead of Dataset-3

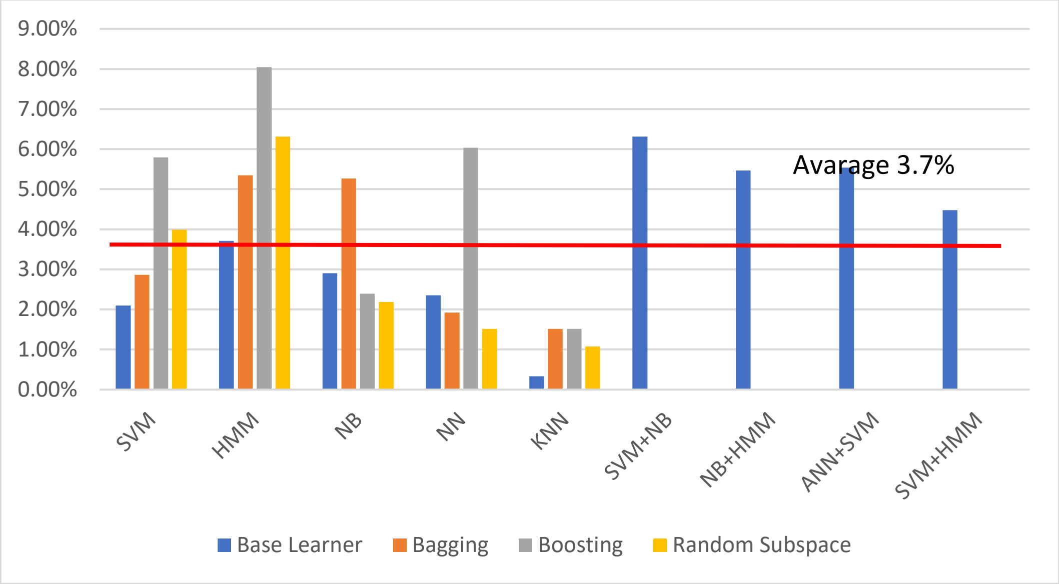


Figure 4.11 The percentage of improvement in classification accuracy F-score that achieved using Dataset-4 instead of Dataset-2

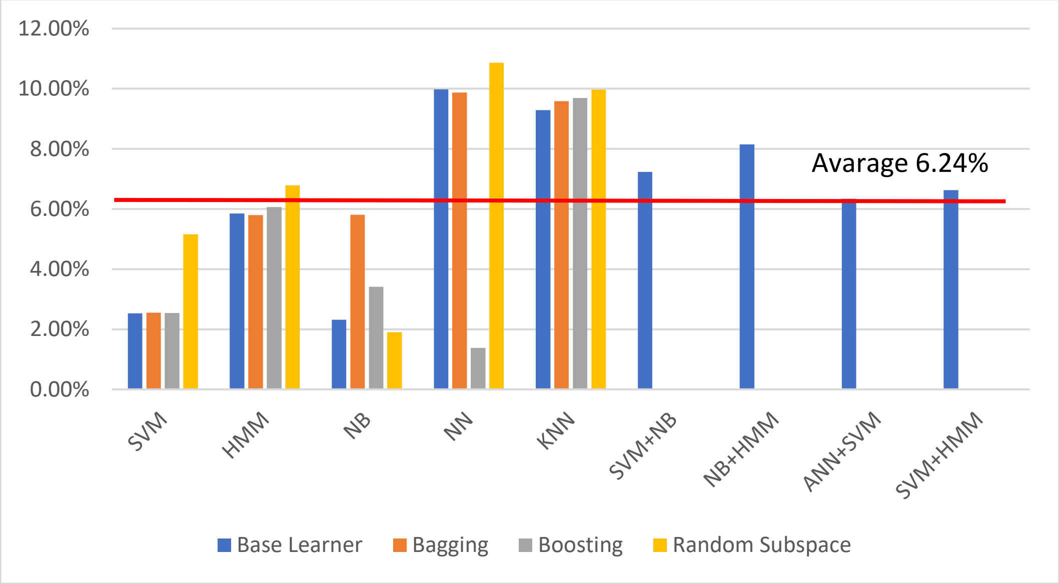


Figure 4.12 The percentage of improvement in classification accuracy F-score that achieved using Dataset-4 instead of Dataset-1

The results in Figures 4.10-4.12 shows that using Dataset-4 has a good impact on improving the classification performance for all classifiers over the other 3 datasets. Using Dataset-4 instead of Dataset-3 provide an average of 1.51% increment in the average F-score of all classifiers and using Dataset-4 instead of Dataset-2 improves the average F-score of all classifiers with a value of 3.7%, while using Dataset-4 instead of Dataset-1 achieved a 6.24% improvement of the average F-score of all classifiers.

4.4.2. Sentiment Regression Model Generation and Testing

To generate our sentiment regression model, we considered testing the commonly used ML regression methods which are SVR MLR MLP. Each regression model was trained and tested individually using our four datasets one at a time. For the evaluation process, 10-fold cross-validation is used to train and test each regression model using the four datasets separately. After applying the test dataset, the MAE and RMSE performance evaluation metrics are obtained for each model as shown in Table 4.9.

Table 4.9 The values MAE and RMSE for each regression model using four different datasets

		Dataset-1	**Dataset-2**	**Dataset-3**	**Dataset-4**
MLP	RMSE	0.25993	0.21628	0.26677	0.25693
	MAE	0.1950	0.16219	0.20008	0.19273
MLR	RMSE	0.24985	0.21439	0.28338	0.27026
	MAE	0.18731	0.16079	0.21251	0.20275
SVR	RMSE	0.18511	**0.18274**	0.24992	0.24519
	MAE	0.13884	**0.13699**	0.18740	0.18392

The results in Tabe 4.9 shows that the SVR regression model has the lowest value of both MAE of 0.13699 and RMSE of 0.18274 over the MLP and MLR model for all the four datasets. The SVR regression model that generated using Dataset-2 (With Bigram features and TF-IDF weights) has the best performance in regard to MAE and RMSE values over all other models. MLP regression model generated using Dataset-3 and Dataset-4 performed better than MLR regression model that generated using the same datasets, however, the MLR models provide lower error than MLP model using Dataset-1 and Dataset-2.

To identify which is the best regression method among (MLR, MLP, and SVR), we calculated the average value of MEA and RMSE for each (MLR, MLP, and SVR) regression method over all our four datasets as shown in Figure 4.13.

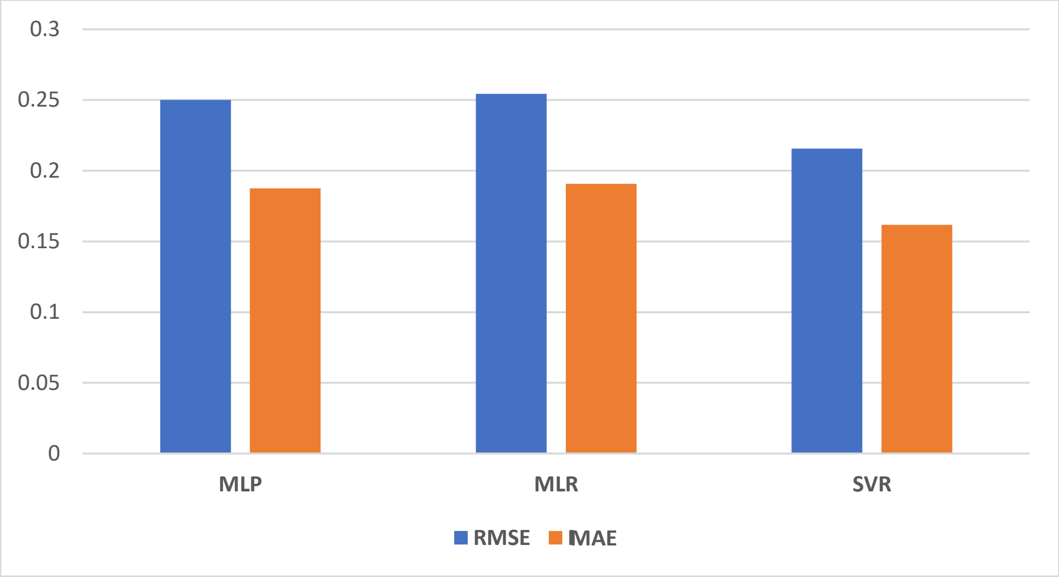


Figure 4.13 The average value of MAE and RMSE for each regression model for all 4 datasets

From the result shown in Figure 4.13, we can conclude that SVR regression method provides the best prediction whereas it has the lowest average value of MAE of 0.16 RMSE of 0.22 over the other MLP and MLR regression methods. The result also shows that the prediction performance of both MLP and MLR is a very similar in the regard to the average value of MEA and RMSE.

<u>Results summary for ML Approach for Document-level Arabic SA</u>

From the experiments result, we can conclude that:

- The datasets with Bigrams features produce classification models with higher accuracy than the datasets with Unigrams features.
- Using TF-IDF rather than TF feature weighting can provide an enhancement in classification performance.

- The maximum classification performance is provided by SVM base learner classifier (which has been proven by many of previous research, that SVM has the more powerful competitiveness in Text classification application especially sentiment classification [15] [16] [17] [15] [27] [28] [40] [41] [47] [48] [49] [50] [51] [52] [56] [57]). In general, SA can be considered as text classification problem with linearly separable categories and, since SVM classification always assumes a hyperplane exist between the classes/categories, so it performs better when the classes/categories are linearly separable as in text classification problem. Also, when the number of data dimensions is very high SVM can be superior to the other classification method in performance wise.
- In classification performance wise, after SVM classifier the HMM classification method takes the second place followed by NB, NN, and KNN. In general, classification method such as SVM, HMM, and NN performs better when it deals with higher dimensional data, However, ANN is more prone to suffer from multiple local minima and overfitting issues which can reduce the performance. On the other hand, classification method such as NB and KNN provide a better performance when working with lower dimensional data ([27] [47] [48] [50] [56] [57] [107]).
- KNN classifier achieved the worst classification performance among all other classifiers (KNN classification algorithm uses the Euclidean distance between the data point to classify a new unknown instance and since the datasets used for sentiment classification tends to have higher dimensionality, this distance measure becomes meaningless and can reduce the classification performance in general [41] [48] [56] [118]).
- Another fact that can affect the classification performance is using imbalanced dataset where the numbers of (positive, negative and neutral) samples that used to train the ML model are not equal.
- The results show the effectiveness of using ensemble learning methods (Bagging, Boosting, and Random Subspace and staking) in term of improving the classification performance.

- The best classification performance is achieved by using Random subspace ensemble method (which combine similar type of classifier models) and staking classifier fusion method (which combine different type of classifier models).
- Using classifier model fusion by stacking method is able to improve the performance in term of accuracy of the all combined (fused) classification models rather than using each single classifier model separately.
- For sentiment score prediction, SVR regression achieved the best performance of sentiment score prediction with the lowest error over the other MLP and MLR regression methods [119] [120] [121] [122]. Generally, SVR regression method does not suffer from the curse of dimensionality problem and work better than the other regression method when dealing with high dimensional data [123] [124].
- Similar to the classification performance results the regression performance is also increased when using Bigrams rather than Unigrams features. However, using TF-IDF feature weighting instead of the TF does not impact the prediction performance of the regression models.
- MLP and MLR regression methods both can learn a linear prediction function, so that these methods can perform comparably in regard to prediction performance. However, MLP method is superior to MLR, where it can learn a nonlinear prediction function also so that it may perform better than MLR [122] [125] [126].
- The maximum classification performance is achieved by using SVM+HMM classifier fusion model with the highest F-score value of 92.35% so that this method is considered for a generation the sentiment classification model that used in our proposed document-level Arabic SA system.
- The maximum prediction performance is achieved by using SVR regression method with the lowest values MAE of 0.13699 and RMSE of 0.18274 so that this method is considered for a generation the sentiment regression (sentiment score prediction) model that used in our proposed document-level Arabic SA system.

4.5. Concept-based Approach for Arabic Sentence-level Sentiment Analysis Evaluation

In this section, we performed comprehensive comparative experiments for evaluating our proposed concept-based Sentence-level Arabic SA system which is based on ML classification and regression approaches. Our goal with this evaluation experiments is to determine and specify the best performing ML classification and regression models to use them for the proposed SA system.

4.5.1. Measure the Coverage of the Translated Arabic SenticNet

In order to evaluate the quality of our translated Arabic SenticNet concept-based sentiment lexicon, we calculate its coverage over our GLASC corpus. The coverage of Ar-SenticNet can be calculated as following;

$$\boldsymbol{Covarage} = \frac{\left[[\boldsymbol{Arabic\ SenticNet}] \cap\ \boldsymbol{GLASC}\right]}{|\boldsymbol{GLASC}|} \tag{4.10}$$

In another word;

$$\boldsymbol{Covarage} = \frac{\boldsymbol{Total\ ferquncy\ of\ commen\ terms\ between\ GLASC\ and\ Ar\ SenticNet}}{\boldsymbol{Total\ frequncy\ of\ terms\ in\ GLASC}} \tag{4.11}$$

Using the formula described above we calculate the coverage of our both Ar-SenticNet_v1 and Ar-SenticNet_v2 which are generated using the process described in Section 3.3. The Ar-SenticNet_v1 contains a total number of concepts of 38,032 where the Ar-SenticNet_v2 contains 48,343 concepts. Based on this coverage calculation formula, the Ar-SenticNet_v1 has obtained a 52.2% coverage over our GLASC corpus, while the Ar-SenticNet_v2 has obtained a 73.3% coverage over our GLASC corpus as shown in Figure 4.14.

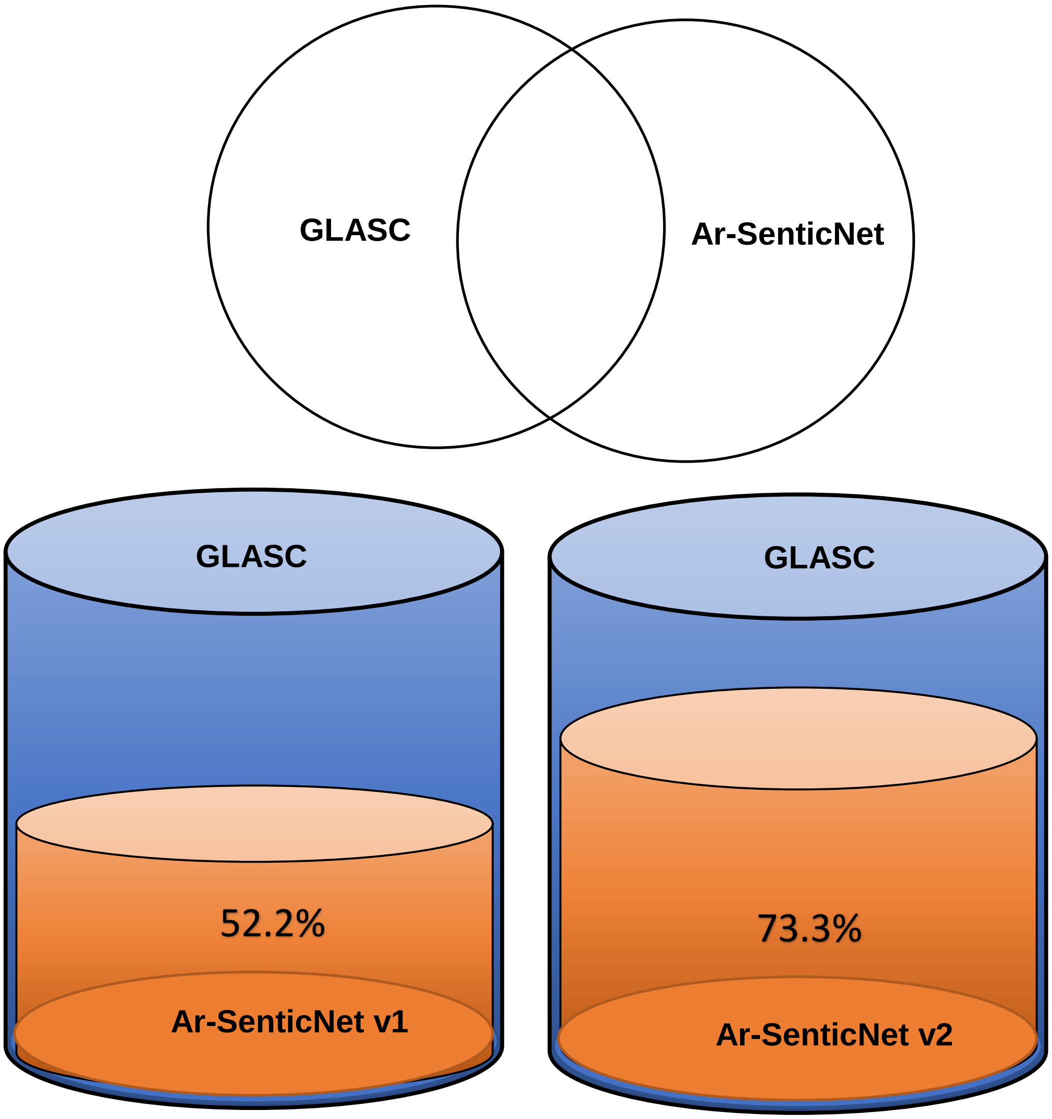


Figure 4.14 A comparison between the coverage of each Ar-SenticNet_v1 and Ar-SenticNet_v2 over our GLASC corpus

The coverage results show that the Ar-SenticNet_v2 has achieved a 21.1% higher coverage than the coverage that achieved by Ar-SenticNet_v1, although the Ar-SenticNet_v2 contains 27.11% more concepts than Ar-SenticNet_v1. The results also show that the Ar-SenticNet_v2 has obtained a sufficient coverage which can demonstrate its effectiveness by covering a wide range of Arabic concepts.

4.5.2. Dataset and Feature Exaction

To build the ML classification and regression models for our proposed concept-based SA system, a dataset is generated from our GLASC corpus. This dataset consists of a total of 1750 sentence organized as (594 negative, 585 positive and 571 neutral) and each sentence contains an average number of 30 words. We applied the feature extraction methods explained in Section 3.4.2, on this dataset to generate different feature vectors which are then used to train and evaluate the system. These different features such as concept-based features CBF, Lexicon based features, Bag of Word features and Word2Vector features are extracted from the dataset. We intended to use different feature combinations in order to evaluate our system. These combinations of features are shown and described in Table below.

Table 4.10 The proposed feature combinations and its descriptions

Features	**Description**
CBF	Using only Concept-based Features
CBF+LEX	Using Concept-based Features together with the Lexicon Based Features
CBF+W2V	Using Concept-based Features together with the Word Vector Features
CBF+ BoW_Uni	Using Concept-based Features together with the Bag of Words Features that generated using Unigrams and TF-IDF weighting
CBF+ BoW_Bi	Using Concept-based Features together with the Bag of Words Features that generated using Bigrams and TF-IDF weighting

Features	Description
CBF+LEX+BoW_Bi	Using Concept-based Features combined with the Lexicon Based Features and the Bag of Words Features that generated using Bigrams and TF-IDF weighting.
CBF+LEX+W2V	Using Concept-based Features combined with the Lexicon Based Features and the Word Vector Features

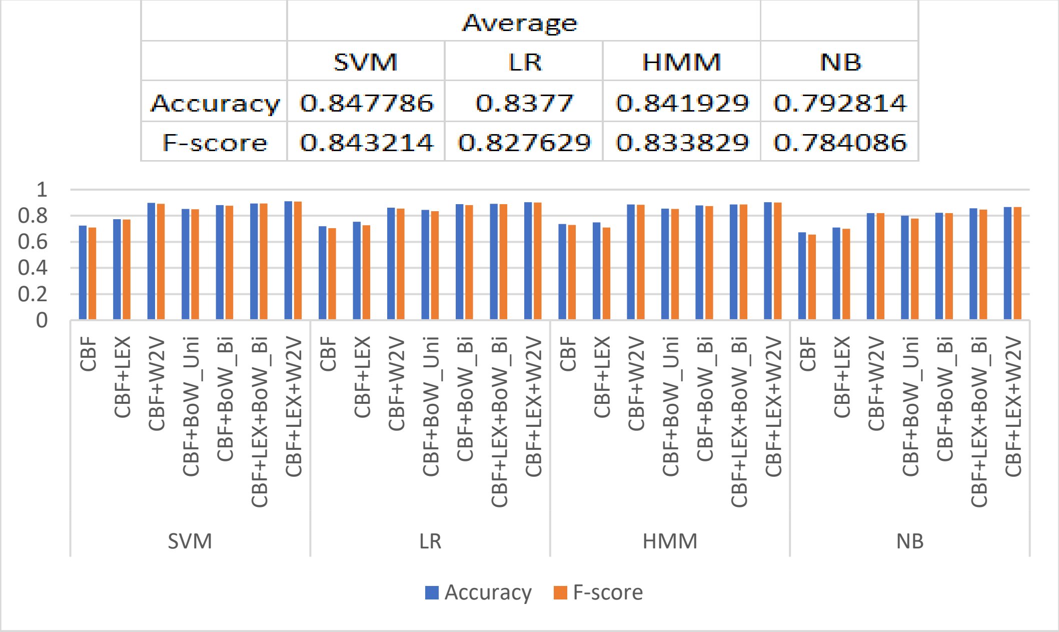
4.5.3. Classification Model Generation and Evaluation

4.5.3.1. Experiments with Base Learners

In order to generate our classification model, we examined four different ML classifier algorithms such as SVM, LR, HMM, and NB. We performed a 10-fold cross validation separately on these four classifiers using different sets of feature combinations and calculated the accuracy and F-score classification model performance evaluation metrics. The obtained accuracy and F-score results for each classifier using distinctive features combinations are reported in Table 4.11 and the average values of classification accuracy and F-score for base learners over all the different features combinations are shown in Figure 4.15.

Table 4.11 The classification Accuracy (A) and F-score (F) for base learners using different features combinations

Features		**SVM**	**LR**	**HMM**	**NB**
CBF	A	0.7245	0.7188	0.7375	0.6734
	F	0.7099	0.7058	0.7284	0.6561
CBF+LEX	A	0.7735	0.7551	0.7478	0.7088
	F	0.7722	0.7274	0.7102	0.6996
CBF+W2V	A	0.8979	0.8607	0.8852	0.8206
	F	0.8924	0.8533	0.8849	0.8200
CBF+BoW_Uni	A	0.8525	0.8457	0.8547	0.8013
	F	0.8491	0.8352	0.8527	0.7787
CBF+BoW_Bi	A	0.8827	0.8880	0.8785	0.8220
	F	0.8771	0.8822	0.8738	0.8210
CBF+LEX+BoW_Bi	A	0.8930	0.8921	0.8855	0.8562
	F	0.8929	0.8895	0.8853	0.8462
CBF+LEX+W2V	A	**0.9104**	0.9035	0.9043	0.8674
	F	**0.9089**	0.9000	0.9015	0.8670



	Average			
	SVM	LR	HMM	NB
Accuracy	0.847786	0.8377	0.841929	0.792814
F-score	0.843214	0.827629	0.833829	0.784086

Figure 4.15 The average values of classification accuracy and F-score of base learners for different features combinations

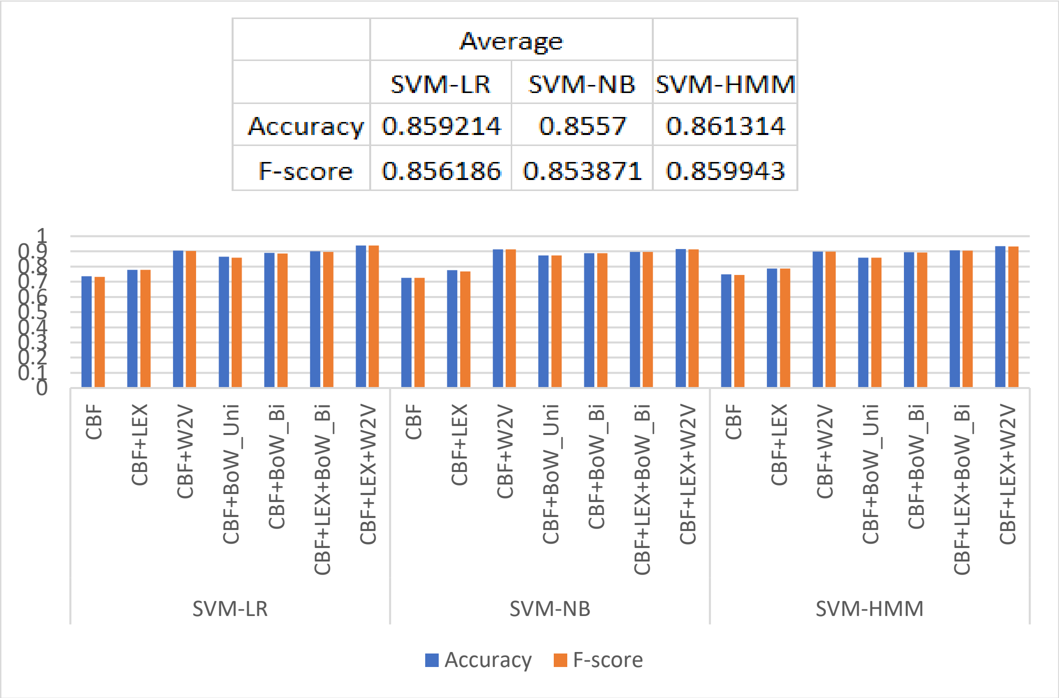
From the results in Table 4.11, it can be shown that the CBF+LEX+W2V features combination provides the best classification performance of all of the classification models that used. SVM classifier provided its best performance of 90.89% F-score using CBF+LEX+W2V features combinations over all other classifiers and other features combinations sets. CBF+LEX+BoW_Bi features combination also provides a very good classification performance for all classifier model less than about only 1.5% form the maximum performance that provided by CBF+LEX+W2V features combination. By using only CBF the best result is obtained by HMM classifier with 72.84% of F-score. Combining CBF with other features such as LEX, BoW, and W2V increases the classification performance for all classifier models used. The result in Figure 4.15 shows that the SVM classification method has the best performance over the other HMM, LR and NB classifiers when it achieved an average value of F-score of 84.32% for all (CBF, CBF+LEX, CBF+W2V, CBF+BoW_Uni, CBF+BoW_Bi, CBF+LEX+BoW_Bi, and CBF+LEX+W2V) features combination.

4.5.3.2. Experiments with Classifier Model Fusion

In order to enhance and improve the classification performance of our classification model we also considered using classifier model fusion approach. We used a different combination of classifiers methods such as (SVM+LR, SVM+NB, and SVM+HMM) as level-0 models. For the level-1 model, we used a multilayer perceptron MLP as meta-classifier to combine the decisions of the level-0 classifier models. Each one of these fused classifier models was individually evaluated using 10-fold cross-validation based on different features sets and the obtained values of classification accuracy and F-score is reported in Table 4.12 below. The average values of classification accuracy and F-score for combined learner over all the different features combinations are shown in Figure 4.16.

Table 4.12 The classification Accuracy (A) and F-score (F) for combined learner using stacking method by using different features combinations

Features		SVM-LR	SVM-NB	SVM-HMM
CBF	A	0.7356	0.7264	0.7482
	F	0.7327	0.7264	0.7456
CBF+LEX	A	0.7784	0.7755	0.7871
	F	0.7783	0.7682	0.7862
CBF+W2V	A	0.9042	0.9134	0.8993
	F	0.9022	0.9126	0.8991
CBF+BoW_Uni	A	0.8651	0.8739	0.8592
	F	0.8582	0.8737	0.8591
CBF+BoW_Bi	A	0.8911	0.8878	0.8935
	F	0.8859	0.8872	0.8914
CBF+LEX+BoW_Bi	A	0.9005	0.8965	0.9079
	F	0.8968	0.8959	0.9059
CBF+LEX+W2V	A	**0.9396**	0.9164	**0.9340**
	F	**0.9392**	0.9131	**0.9323**



	Average		
	SVM-LR	SVM-NB	SVM-HMM
Accuracy	0.859214	0.8557	0.861314
F-score	0.856186	0.853871	0.859943

Figure 4.16 The average classification accuracy and F-score of combined learner using stacking method for all different features combinations

When comparing the results of combined classifier models with the result of base learners, it’s shown that the combined models have an impact on increasing the classification performance. Similar to based learner results the best classification performance is obtained using CBF+LEX+W2V features combinations. for all used features combination, the best performance is achieved by SVM-LR combined classification. SVM-LR model proved its best classification performance of 93.92% F-score using CBF+LEX+W2V features combinations. The best performance using only CBF is achieved by SVM-HMM model with F-score of 74.56%. using CBF combined with the other features such as LEX, BoW, and W2V increases the classification performance for all classifier models used. The result in Figure shows that the SVM-HMM fused classification model has the best performance over the other combined classifiers where its achieved an average value of F-score of 85.99% for all features combination.

In order to identify the impact of classifier combination instead of using individual classifiers, in improving classification performance, we calculate percentage value of the improvement in classification accuracy and the F-score using the combined classifiers instead of using each classifier individually. The results are shown in Figure 4.17.

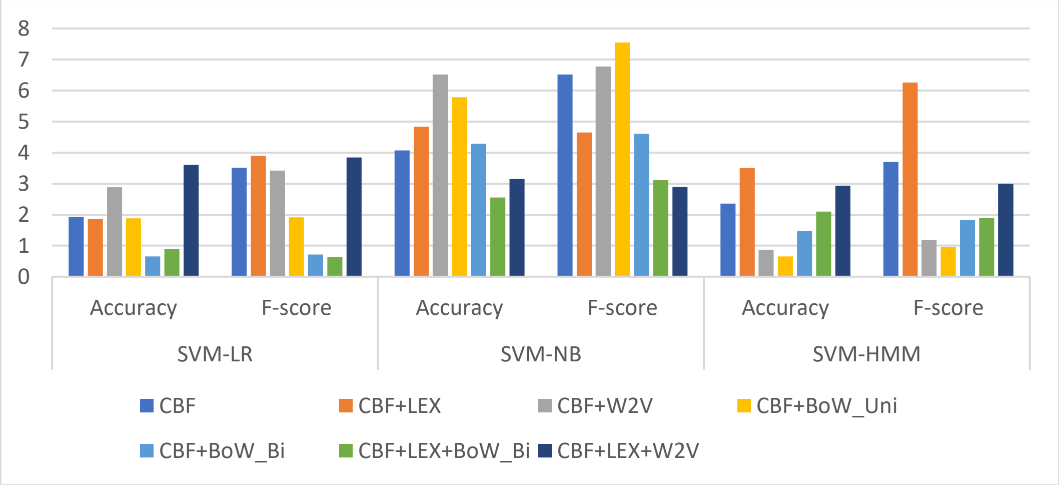


Figure 4.17 The percentage of improvement in the average of classification accuracy and F-score for each combined learner using stacking method for different features combinations

The result in Figure 4.17 shows that using (SVM+NB) classifiers combination is provide a maximum value of 7.55% improvement in F-score over the performance achieved by (SVM and NB) classifiers individually with CF+BoW_Uni features set, and an average improvement of 5.15% for all different features combinations. This means that the stronger SVM classifier is capable of enhancing the performance of the weaker NB classifier when it used together. The average improvement in F-score for the all of the different features combinations by using SVM-LR classifier combination is 2.55% and it’s 2.62% when using SVM-HMM classifier combination.

Although the best F-score result is achieved by the SVM, HMM and LR classifiers when it used individually with CBF+LEX+W2V features combinations, there is only 3% improvement in F-score using the (SVM+HMM) combination and 3.84% using the (SVM+LR) combination, with the same CBF+LEX+W2V features combinations.

Since we used different features combinations in order to evaluate our sentiment classification system, we investigate the impact of using different feature combination sets on the classification performance. In order to calculate the percentage of performance improvement using these different features combinations, since CBF features are the common point in all our features combinations, we first selected the CBF as a baseline for our experiment, then we calculate the percentage of improvement in F-score for each classifier/classifier-combinations, when a different feature is combined with baseline features. The results in Figure 4.18 shows the percentage of improvement in F-score for each classifier when using the other features are combined with CBF baseline features.

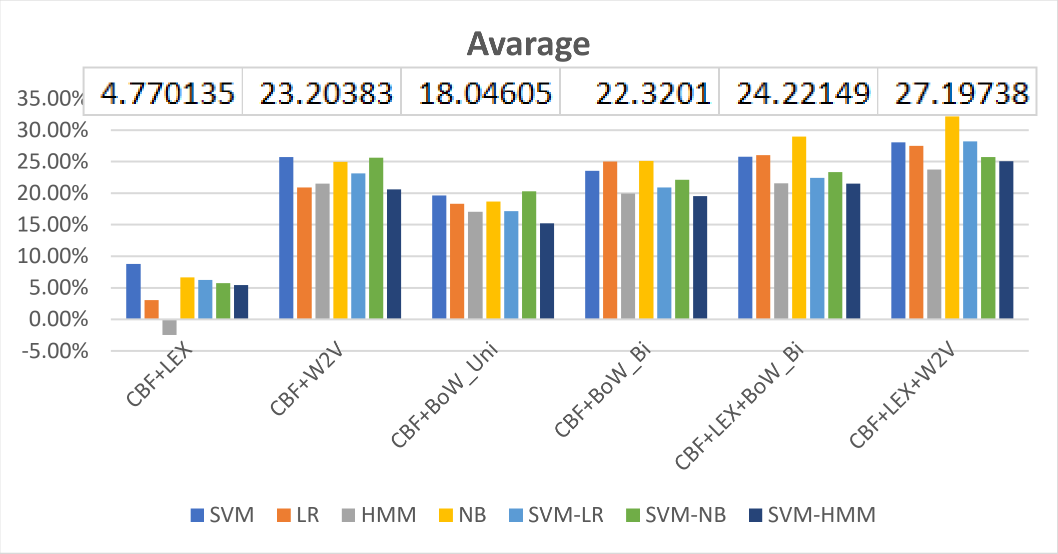


Figure 4.18 The percentage of improvement in F-score for each classifier when using the other features are combined with CBF baseline features

The results in Figure 4.18 shows that the maximum average of improvement in F-score for all used classifiers is 27.2% which is achieved by combining LEX and W2V features with the baseline CBF features. Combining LEX features with the CBF features improved the average F-score of all classifiers by only 4.77%. While combining only W2V features with the CBF is achieved 22.2% improvement in the average F-score of all classifiers. The results also show that combining BoW features with CBF also result in an improvement in the average F-score of all classifiers.

4.5.4. Sentiment Regression Model Generation and Testing

To generate our sentiment regression model, we considered testing the commonly used ML regression methods which are SVR, MLR, and MLP. For the evaluation process, 10-fold cross-validation is used to train and test separately each one of regression model using different features combinations. After applying the test dataset, MAE and RMSE performance evaluation metrics are obtained for each model and reported in Table 4.13 and show in Figure 4.19.

Table 4.13 The values MAE and RMSE for each regression model using different features combinations

Features		SVR	MLR	MLP
CBF	RMSE	0.13494	0.12654	0.13914
	MAE	0.10142	0.09561	0.10459
CBF+LEX	RMSE	0.10195	0.11247	0.11068
	MAE	0.07648	0.08369	0.08282
CBF+W2V	RMSE	**0.07818**	0.08387	0.10203
	MAE	**0.05875**	0.06369	0.07692
CBF+BoW_Uni	RMSE	0.21865	0.19953	0.20798
	MAE	0.16148	0.14948	0.15765
CBF+BoW_Bi	RMSE	0.22170	0.22274	0.22286
	MAE	0.16505	0.16530	0.16665
CBF+LEX+BoW_Bi	RMSE	0.23062	0.22899	0.22916
	MAE	0.17337	0.17242	0.17141
CBF+LEX+W2V	RMSE	0.15716	0.15536	0.17125
	MAE	0.11778	0.11574	0.12841

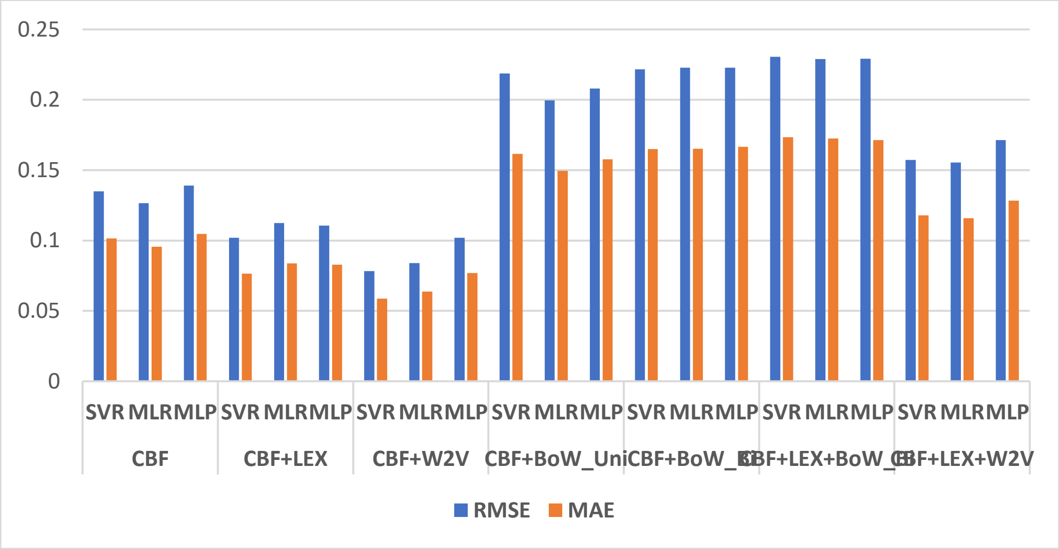
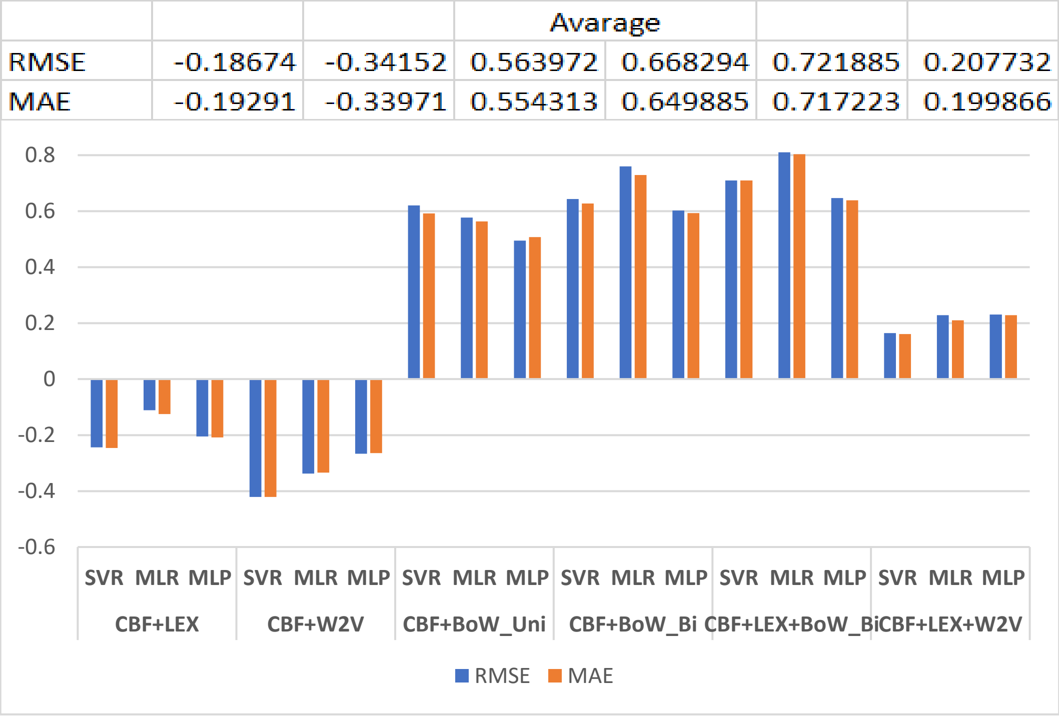


Figure 4.19 The values MAE and RMSE for each regression model using different features combinations

From the results in Figure 4.19, we can conclude that the best prediction error of the SVR, MLR, and MLP are achieved by using CF+W2V features combinations. SVR regression method provides the best prediction by achieving the lowest value of MAE of 0.059 and RMSE of 0.078 over the other SVR and MLR regression methods by using CF+W2V features combinations.

In order to calculate the percentage of performance improvement using these different features combinations, and since CBF features are the common point in all our features combinations, we first selected the CBF as a baseline for our experiment, then we calculate the percentage of improvement in both of RMSE and MAE for each SVR, MLR, and MLP when a different feature is combined with baseline features. The results in Figure 4.20 shows the percentage of improvement in RMSE and MAE for each regression method, when the other features are combined with CBF baseline features.



			Avarage			
RMSE	-0.18674	-0.34152	0.563972	0.668294	0.721885	0.207732
MAE	-0.19291	-0.33971	0.554313	0.649885	0.717223	0.199866

Figure 4.20 The percentage of improvement in RMSE and MAE for each regression method by combining other features with CBF baseline features

From the result shown in Figure 4.20, for all regression methods combining W2V features to the CBF baseline features results in an improvement in the average prediction performance for all regression method used by reducing the average RMSE and MEA by 34% compared to the prediction performance achieved by only the baseline CBF features. Combining LEX features with the baseline CBF features is also improved the prediction performance for all regression method used when it reduced the average RMSE and MEA by 19%. Combining BoW features with the baseline CBF features produced a negative impact on the average prediction performance for all regression method by increasing the average value of both RMSE and MEA.

Results summary for Concept-based Approach for Arabic Sentence-level SA

From the experiments result, we can conclude that:

- Using the baseline CBF features individually results in a poor classification performance for all used classification models.
- Combining various features such as (LEX, BoW_Uni, Bow_Bi, W2V) with the baseline CBF features results in improving the classification performance for all used classification models.
- The maximum improvement in classification performance for all used classifiers features is achieved using (CBF+LEX+W2V) features combinations.
- Combining LEX features with the CBF+W2V and CBF+BoW features combinations resulted in an improvement in classification performance for all used classifiers. However, combining only LEX features with the CBF features resulted in a less improvement in classification performance compared to the improvement achieved by combining only W2V or BoW features with the CBF features.
- Combining BoW based features with the baseline CBF features is also achieved a responsible improvement in classification performance.
- The maximum classification performance is achieved by SVM base learner classifier for the all different features combinations followed by HMM then LR and NN classifier.
- Using classifier model fusion by stacking method is able to improve the accuracy of the all combined (fused) classification models rather than using each single classifier model separately.
- For sentiment score prediction, the best prediction error results from the SVR, MLR and MLP regression methods are achieved by using CF+W2V features combinations.

- SVR regression method provides the best prediction by achieving the lowest value of MAE and RMSE over the other SVR and MLR regression methods with CF+W2V features combinations.
- Using W2V or LEX features combined with the baseline CBF features results in improving the prediction performance by reducing RMSE and MEA values of the regression models. However, combining BoW features with the baseline CBF features produced a negative impact on the average prediction performance for all regression method by increasing the average value of both RMSE and MEA.
- The maximum classification performance is achieved by using SVM+LR classifier fusion model with the highest F-score value of 93.92%, so that this method is considered for the generation the sentiment classification model that used in our proposed concept-based Sentence-level Arabic SA system.
- The maximum prediction performance is achieved by using SVR regression method with the lowest RMSE and MAE values of 0.078 and 0.059, so that this method is considered for generating the sentiment regression (sentiment score prediction) model that used in our proposed concept-based Sentence-level Arabic SA system.

5. CONCLUSİON

In this thesis, we presented a large-scale Arabic SA corpus called GLASC, which built using online Arabic news articles and metadata provided by the Bigdata resource GDELT. Our corpus consists of a total of 620,082 Arabic news articles divided into three categories (Positive, Negative and Neutral). Besides that, our corpus also provides a sentiment rating by assigning a sentiment score in a range between -1 and 1 for each article.

We carried out two different types of experiments in order to evaluate the quality of the generated GLASC corpus. The first evaluation experiment involves using statistical measures to calculate the percentage of Positive, Negative and Neutral terms in each Positive, Negative and Neutral category in our corpus based on Arabic sentiment lexicon called (ArSenL). The second evaluation experiment involves comparing the term rank to term frequency distribution of our GLASC corpus to the ideal Zipf distribution. These evaluation experiments confirmed the quality of our GLASC corpus when its Positive, Negative and Neutral categories contain the proper ratio of (Positive, Negative and Neutral) terms, and the term rank to term frequency distribution of the corpus fitted very well to the Zipf-Mandelbrot law distribution.

To our best knowledge, this corpus can be considered as the largest resource available for Arabic language and we believe it will provide a significant contribution not only to SA but also to a wide range of Arabic NLP applications in Big Data domain.

We used our GLASC corpus to build an Arabic document-level SA system based on ML classification and regression approaches, when a ML-based classifier model is used for assigning an Arabic document into one of three various categories (Positive, Negative or Neutral) and a ML-based regression model used for predicting the sentiment score of the Arabic document based on its sentiment orientation.

For training the sentiment classifier and regression models, we generated four datasets from our corpus using different feature extraction and feature weighting methods. We performed a comparative study, involving testing a wide range of, classification methods such as (SVM, HMM, NB, NN, and KNN) and regression methods such as (SVR, MLR and MLP) which are commonly used for sentiment analysis task. For the performance evaluation of our ML approaches, we considered the accuracy and F-score metrics for evaluating the classification models and MAE and RMSE for evaluating the regression models.

In addition to that we also investigated several types of ensemble learning methods such as (“Bagging”, “Boosting”, “Random subspace” and “Staking’) to verify its impact on improving the classification performance for sentiment analysis, using different comprehensive empirical experiments.

Our experiments show that the best classification performance is achieved using a dataset with Bigram features and TF-IDF weights over the other three datasets.

The obtained results showed that as a base learner SVM and HMM have achieved the best results with a value of F-score of 87.06% and 86.75% respectively.

Our experiments result also verified the impact of using ensemble learning methods (“Bagging”, “Boosting”, ‘Random Subspace” and “Staking’) in term of improving the classification performance.

The ensemble model of SVM using Random Subspace method has achieved the best classification accuracy of 90.21% of an F-score and the ensemble model of HMM using the same Random Subspace method has achieved an F-score of 89.21%.

Regarding to the results of our experiments, the maximum classification performance is achieved by using stacking classifier fusion method with the highest value of 92.35% of F-score for the SVM+HMM classifiers fusion and a value of 91.42% of F-score for the SVM+NB classifiers fusion.

For the ML-based regression model, our experimental results show that SVR regression model which is generated by using the Dataset-2 (With Bigram features and TF-IDF weights) provided the best prediction performance by achieving the lowest value of MAE of 0.13699 and RMSE 0.18274.

A concept based sentiment lexicon for the Arabic language is generated by translating the English version of the concept-based sentiment lexicon SenticNet to Arabic using two-way translation and extension process. The translation process utilizes the English-Arabic cross-language mapping which provided by WordNet and the Google translation service. Then the translated Arabic concept based sentiment lexicon is extended by adding more senses which are obtained from Arabic WordNet.

We applied this translation and extension approach on both English SenticNet_v3 and the recently released SenticNet_v4 to generate two different Arabic versions of SenticNet. Firstly, translation and extension process has applied on 33k concept English SenticNet_v3 and resulted in the Ar-SenticNet_v1 that contain 31k Arabic concepts, then the same process has again applied on the 50k concept English SenticNet_v4 and resulted in the Ar-SenticNet_v2 with 48k Arabic concepts.

In order to evaluate the quality of our translated Arabic SenticNet concept based sentiment lexicons, we calculate its coverage over our GLASC corpus based on coverage calculation formula that described in Section 4.6.1. based on the performed calculation, the Ar-SenticNet_v1 has obtained a 52.2% coverage over our GLASC corpus where the Ar-SenticNet_v2 has obtained a 73.3% coverage over our GLASC corpus. This means that the Ar-SenticNet_v2 provide a very good cover to most of the concepts that found in our large-scale GLASC corpus.

We also build a concept based SA system for Arabic Sentence-level sentiment analysis using our previously mentioned Ar-SenticNet concept-based sentiment lexicon and a variety of ML approaches. For extracting the concepts from the Arabic sentence, we proposed and performed a rule-based concept extraction algorithm called semantic parser. In order to generate the candidate concepts list for an Arabic sentence, this semantic parser utilizes a variety of freely available grammatical and morphological analysis tools for the Arabic language beside to the grammatical rules of the Arabic concepts.

We also presented and used different types of feature extraction and representation techniques for building the concept-based Sentence-level Arabic SA system. These techniques are used to extract various feature sets from the input sentence, which are used to build the ML decision model. These feature sets are concept based features CBF, lexicon based features LEX, Bag of Word features BoW and Word2Vector features W2V.

For building the ML-based decision model used in our concept-based Sentence-level Arabic SA system we used different types of ML classification methods such as (SVM, HMM, NB and LR) and different types of ML-based regression methods such as (MLR, MLP, and SVR). In order to improve the classification performance, we also used classifier fusion method for combining classification models such as (SVM-HMM, SVM-NB, and SVM-LR). For training these ML-based models we generated a sentence based dataset form our GLASC corpus and carried out a comprehensive and comparative experiments using different combinations of the feature sets that mentioned earlier with the baseline concept based features CBF. The features combinations that we used are (CBF, CBF+LEX, CBF+W2V, CBF+BoW_Uni, CBF+ BoW_Bi, CBF+LEX+BoW_Bi and CBF+LEX+W2V).

Our experiment results show that the best performance for the classification model is achieved by using SVM classifier which has obtained an F-score value of 90.89% using CBF+LEX+W2V features combinations, while the combined SVM-LR model has obtained a better classification performance of 93.23% F-score using the same CBF+LEX+W2V features combinations. For the ML-based regression model, our experimental results show that SVR regression method provides the best prediction by achieving the lowest value of MAE of 0.059 and RMSE of 0.078 using CF+W2V features combinations.

For the future works; we are considering using an approach similar to the one used in [127] for expanding both of the Arabic sentiment lexicon SentiWordNet (ArSenL) and the Arabic concept-based sentiment lexicon SenticNet, using our large-scale corpus GLASC. We are also considering using rule-based SA approaches together with concept-based SA approaches, which can lead to increase in the precision and accuracy of SA by using a language dependent grammatical rules.

REFERENCES

[1] T. Nasukawa and J. Yi, “Sentiment Analysis: Capturing Favorability Using Natural Language Processing,” in Proceedings of the 2Nd International Conference on Knowledge Capture, New York, NY, USA, **2003**, pp. 70–77.

[2] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of the 12th international conference on World Wide Web, **2003**, pp. 519–528.

[3] C. Elliot, “The affective reasoner: A process model of emotions in a multi-agent system. **1992**,” Northwest. Univ. Inst. Learn. Sci. Northwest. IL, vol. 48.

[4] A. Ortony and T. J. Turner, “What’s basic about basic emotions?,” Psychol. Rev., vol. 97, no. 3, p. 315, **1990**.

[5] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends® Inf. Retr., vol. 2, no. 1–2, pp. 1–135, **2008**.

[6] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” Comput. Linguist., vol. 37, no. 2, pp. 267–307, **2011**.

[7] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, **2002**, pp. 79–86.

[8] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, **2014**.

[9] S.-M. Kim and E. Hovy, “Determining the sentiment of opinions,” in Proceedings of the 20th international conference on Computational Linguistics, **2004**, p. 1367.

[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of twitter data,” in Proceedings of the workshop on languages in social media, **2011**, pp. 30–38.

[11] Y. Wilks and M. Stevenson, “The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation,” Nat. Lang. Eng., vol. 4, no. 2, pp. 135–143, **1998**.

[12] A. Aizawa, “The feature quantity: an information theoretic perspective of tfidf-like measures,” in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, **2000**, pp. 104–111.

[13] T. Scheffer and S. Wrobel, “Text Classification beyond...,” in Proceedings of the ICML-Workshop on Text Learning, **2002**.

[14] J. I. Serrano and M. D. del Castillo, “Text representation by a computational model of reading,” in International Conference on Neural Information Processing, **2006**, pp. 237–246.

[15] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” Mach. Learn. ECML-98, pp. 137–142, **1998**.

[16] V. N. Vapnik and V. Vapnik, Statistical learning theory, vol. 1. Wiley New York, **1998**.

[17] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: tasks, approaches and applications,” Knowl.-Based Syst., vol. 89, pp. 14–46, **2015**.

[18] A. Esuli and F. Sebastiani, “SentiWordNet: a high-coverage lexical resource for opinion mining,” Evaluation, pp. 1–26, **2007**.

[19] A. Hamouda and M. Rohaim, “Reviews classification using sentiwordnet lexicon,” in World congress on computer science and information technology, **2011**.

[20] F.-R. Chaumartin, “UPAR7: A knowledge-based system for headline sentiment tagging,” in Proceedings of the 4th International Workshop on Semantic Evaluations, **2007**, pp. 422–425.

[21] K. Denecke, “Using sentiwordnet for multilingual sentiment analysis,” in Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on, **2008**, pp. 507–512.

[22] M. Ptaszynski, R. Rzepka, S. Oyama, M. Kurihara, and K. Araki, “A Survey on Large Scale Corpora and Emotion Corpora,” Inf. Media Technol., vol. 9, no. 4, pp. 429–445, **2014**.

[23] M. Ptaszynski, P. Dybala, R. Rzepka, K. Araki, and Y. Momouchi, “YACIS: A five-billion-word corpus of Japanese blogs fully annotated with syntactic and affective information,” in Proceedings of The AISB/IACAP World Congress, **2012**, pp. 40–49.

[24] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” Int. J. Inf. Manag., vol. 35, no. 2, pp. 137–144, **2015**.

[25] R. H. Baayen, Word frequency distributions, vol. 18. Springer Science & Business Media, **2001**.

[26] G. K. Zipf, “The psycho-biology of language.,” **1935**.

[27] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, “OCA: Opinion corpus for Arabic,” J. Assoc. Inf. Sci. Technol., vol. 62, no. 10, pp. 2045–2054, **2011**.

[28] M. Abdul-Mageed and M. T. Diab, “AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.,” in LREC, **2012**, pp. 3907–3914.

[29] M. A. Aly and A. F. Atiya, “LABR: A Large Scale Arabic Book Reviews Dataset.,” in ACL (2), **2013**, pp. 494–498.

[30] M. Abdul-Mageed, M. Diab, and S. Kübler, “SAMAR: Subjectivity and sentiment analysis for Arabic social media,” Comput. Speech Lang., vol. 28, no. 1, pp. 20–37, Jan. **2014**.

[31] M. Al-Smadi, O. Qawasmeh, B. Talafha, and M. Quwaider, “Human annotated arabic dataset of book reviews for aspect based sentiment analysis,” in Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on, **2015**, pp. 726–730.

[32] H. ElSahar and S. R. El-Beltagy, “Building Large Arabic Multi-domain Resources for Sentiment Analysis.,” in CICLing (2), **2015**, pp. 23–34.

[33] “The GDELT Project.” [Online]. Available: https://www.gdeltproject.org/. [**Accessed: 08-Nov-2017**].

[34] E. Cambria and B. White, “Jumping NLP curves: A review of natural language processing research,” IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48–57, **2014**.

[35] B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain, “Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach,” Cogn. Comput., vol. 7, no. 4, pp. 487–499, **2015**.

[36] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. "Introduction to WordNet: An on-line lexical database." International journal of lexicography 3, no. 4, pp. 235-244,**1990**.

[37] E. Cambria, R. Speer, C. Havasi, and A. Hussain, “SenticNet: A Publicly Available Semantic Resource for Opinion Mining.,” presented at the AAAI fall symposium: commonsense knowledge, **2010**, vol. 10.

[38] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, “Knowledge-based approaches to concept-level sentiment analysis,” IEEE Intell. Syst., vol. 28, no. 2, pp. 12–14, **2013**.

[39] M. Korayem, D. J. Crandall, and M. Abdul-Mageed, “Subjectivity and Sentiment Analysis of Arabic: A Survey.,” in AMLTA, **2012**, pp. 128–139.

[40] N. Farra, E. Challita, R. A. Assi, and H. Hajj, “Sentence-Level and Document-Level Sentiment Mining for Arabic Texts,” in Proceedings of the 2010 IEEE International Conference on Data Mining Workshops, Washington, DC, USA, **2010**, pp. 1114–1119.

[41] N. Abdulla, N. Mahyoub, M. Shehab, and M. Al-Ayyoub, “Arabic sentiment analysis: Corpus-based and lexicon-based,” in Proceedings of The IEEE conference on Applied Electrical Engineering and Computing Technologies (AEECT), **2013**.

[42] A. Montoyo, P. MartíNez-Barco, and A. Balahur, Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Elsevier, **2012**.

[43] “Top 10 Languages Used on the Internet,” Accredited Language Services. [Online]. Available: https://www.accreditedlanguage.com/2016/09/13/top-10-languages-used-on-the-internet. **[Accessed: 08-Nov-2017].**

[44] A. Abdelali, J. Cowie, and H. S. Soliman, “Arabic information retrieval perspectives,” in Proceedings of the 11th Conference on Natural Language Processing, Journes d’Etude sur la Parole-Traitement Automatique des Langues Naturelles (JEP-TALN), **2004**, pp. 391–400.

[45] “Arabic - Wikipedia.” [Online]. Available: https://en.wikipedia.org/wiki/Arabic. **[Accessed: 08-Nov-2017]**.

[46] M. Biltawi, W. Etaiwi, S. Tedmori, A. Hudaib, and A. Awajan, “Sentiment classification techniques for Arabic language: A survey,” in Information and Communication Systems (ICICS), 2016 7th International Conference on, **2016**, pp. 339–346.

[47] A. Shoukry and A. Rafea, “Sentence-level Arabic sentiment analysis,” in Collaboration Technologies and Systems (CTS), 2012 International Conference on, **2012**, pp. 546–550.

[48] A. Mountassir, H. Benbrahim, and I. Berrada, “Some methods to address the problem of unbalanced sentiment classification in an arabic context,” in Information Science and Technology (CIST), 2012 Colloquium in, **2012**, pp. 43–48.

[49] S. Ahmed, M. Pasquier, and G. Qadah, “Key issues in conducting sentiment analysis on Arabic social media text,” in 2013 9th International Conference on Innovations in Information Technology (IIT), **2013**, pp. 72–77.

[50] N. A. Abdulla, M. Al- Mountassirassir, and M. N. Al-Kabi, “An extended analytical study of arabic sentiments,” Int. J. Big Data Intell. 1, vol. 1, no. 1–2, pp. 103–113, **2014**.

[51] M. Elmasry, T. Soliman, and A.-R. Hedar, “Sentiment Analysis of Arabic Slang Comments on Facebook”, vol. 12. **2014**.

[52] N. El-Makky, N. Khaled, E. Alaa, A. Esraa, H. Omneya, M. Samar, and I. Shimaa. "Sentiment analysis of colloquial Arabic tweets." In ASE BigData/SocialInformatics/PASSAT/BioMedCom 2014 Conference, Harvard University, pp. 1-9. **2014**.

[53] M. Elarnaoty, S. AbdelRahman, and A. Fahmy, “A machine learning approach for opinion holder extraction in Arabic language,” IJAIA International Journal of Artificial Intelligence & Applications, **2012**.

[54] S. R. El-Beltagy and A. Ali, “Open issues in the sentiment analysis of Arabic social media: A case study,” in Innovations in information technology (iit), 2013 9th international conference on, **2013**, pp. 215–220.

[55] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh, and M. Haidar, “An opinion analysis tool for colloquial and standard Arabic,” in The Fourth International Conference on Information and Communication Systems (ICICS 2013), **2013**, pp. 23–25.

[56] R. M. Duwairi, R. Marji, N. Sha’ban, and S. Rushaidat, “Sentiment analysis in arabic tweets,” in Information and communication systems (icics), 2014 5th international conference on, **2014**, pp. 1–6.

[57] R. M. Duwairi, M. Alfaqeh, M. Wardat, and A. Alrabadi, “Sentiment analysis for Arabizi text,” in Information and Communication Systems (ICICS), 2016 7th International Conference on, **2016**, pp. 127–132.

[58] M. Abdul-Mageed and M. T. Diab, “AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis.,” in LREC, **2012**, pp. 3907–3914.

[59] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, “A large scale Arabic sentiment lexicon for Arabic opinion mining,” ANLP 2014, vol. 165, **2014**.

[60] “The GDELT Event Database Data Format Codebook v2.0.”. [Online]. Available: http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf. **[Accessed: 08-Nov-2017]**.

[61] “GDELT Translingual: Translating the Planet,” GDELT Blog, [Online]. Available: https://blog.gdeltproject.org/gdelt-translingual-translating-the-planet/. **[Accessed: 08-Nov-2017]**.

[62] Y. Demchenko, C. De Laat, and P. Membrey, “Defining architecture components of the Big Data Ecosystem,” in Collaboration Technologies and Systems (CTS), 2014 International Conference on, **2014**, pp. 104–112.

[63] D. J. Gerner, P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr, “Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions,” Int. Stud. Assoc. New Orleans, **2002**.

[64] D. J. Bodas-Sagi and J. M. Labeaga, “Using GDELT data to evaluate the confidence on the Spanish Government energy policy,” Int. J. Interact. Multimed. Artif. Intell., vol. 3, no. Special Issue on Big Data and AI, **2016**.

[65] B. Agarwal and N. Mittal, “Machine learning approach for sentiment analysis,” in Prominent feature extraction for sentiment analysis, Springer, **2016**, pp. 21–45.

[66] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Comput Surv, vol. 34, no. 1, pp. 1–47, Mar. **2002**.

[67] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, **1995**.

[68] P. Bermejo, J. A. Gámez, and J. M. Puerta, “Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets,” Expert Syst. Appl., vol. 38, no. 3, pp. 2072–2080, **2011**.

[69] S. Soni and A. Sharaff, “Sentiment Analysis of Customer Reviews Based on Hidden Markov Model,” in Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), New York, NY, USA, **2015**, p. 12:1–12:5.

[70] S. Tan, “An effective refinement strategy for KNN text classifier,” Expert Syst. Appl., vol. 30, no. 2, pp. 290–298, Feb. **2006**.

[71] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” J. Biomed. Inform., vol. 35, no. 5, pp. 352–359, Oct. **2002**.

[72] C. M. Bishop, “Pattern recognition and machine learning.” springer, **2006**.

[73] G. Grégoire, “Multiple Linear Regression,” Eur. Astron. Soc. Publ. Ser., vol. 66, pp. 45–72, Jan. **2014**.

[74] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” Bull. Math. Biophys., vol. 5, no. 4, pp. 115–133, Dec. **1943**.

[75] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, “Sentiment classification: The contribution of ensemble learning,” Decis. Support Syst., vol. 57, pp. 77–93, **2014**.

[76] Y. Su, Y. Zhang, D. Ji, Y. Wang, and H. Wu, “Ensemble Learning for Sentiment Classification,” in Chinese Lexical Semantics, **2012**, pp. 84–93.

[77] D. Ruta and B. Gabrys, “An overview of classifier fusion methods,” Comput. Inf. Syst., vol. 7, no. 1, pp. 1–10, **2000**.

[78] S. Džeroski and B. Ženko, “Is combining classifiers with stacking better than selecting the best one?,” Mach. Learn., vol. 54, no. 3, pp. 255–273, **2004**.

[79] R. Wille, “Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts,” in Ordered Sets, Springer, Dordrecht, **1982**, pp. 445–470.

[80] U. Priss, “Formal concept analysis in information science,” Annu. Rev. Inf. Sci. Technol., vol. 40, no. 1, pp. 521–543, Jan. **2006**.

[81] S.-T. Li and F.-C. Tsai, “A fuzzy conceptualization model for text mining with application in opinion polarity classification,” Knowl.-Based Syst., vol. 39, no. Supplement C, pp. 23–33, Feb. **2013**.

[82] E. Kontopoulos, C. Berberidis, T. Dergiades, and N. Bassiliades, “Ontology-based sentiment analysis of twitter posts,” Expert Syst. Appl., vol. 40, no. 10, pp. 4065–4074, Aug. **2013**.

[83] A. Mudinas, D. Zhang, and M. Levene, “Combining Lexicon and Learning Based Approaches for Concept-level Sentiment Analysis,” in Proceedings of the First

International Workshop on Issues of Sentiment Discovery and Opinion Mining, New York, NY, USA, **2012**, p. 5:1–5:8.

[84] E. Cambria, C. Havasi, and A. Hussain, "SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis," Proc. 25th Int. Fla. Artif. Intell. Res. Soc. Conf. FLAIRS-25, pp. 202–207, Jan. **2012**.

[85] E. Cambria, S. Poria, R. Bajpai, and B. W. Schuller, "SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives." In COLING, pp. 2666-2677. **2016**.

[86] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, "Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis," Inf. Process. Manag., vol. 52, no. 1, pp. 36–45, Jan. **2016**.

[87] S. Poria, A. Gelbukh, E. Cambria, D. Das, and S. Bandyopadhyay, "Enriching SenticNet Polarity Scores through Semi-Supervised Fuzzy Clustering," in 2012 IEEE 12th International Conference on Data Mining Workshops, **2012**, pp. 709–716.

[88] A. Qazi, R. G. Raj, M. Tahir, E. Cambria, and K. B. S. Syed, "Enhancing Business Intelligence by Means of Suggestive Reviews," The Scientific World Journal, **2014**.

[89] M. Dragoni, A. G. B. Tettamanzi, and C. da C. Pereira, "A Fuzzy System for Concept-Level Sentiment Analysis," in Semantic Web Evaluation Challenge, **2014**, pp. 21–27.

[90] M. Araújo, P. Gonçalves, M. Cha, and F. Benevenuto, "iFeel: A System That Compares and Combines Sentiment Analysis Methods," in Proceedings of the 23rd International Conference on World Wide Web, New York, NY, USA, **2014**, pp. 75–78.

[91] H. H. Wu, A. C. R. Tsai, R. T. H. Tsai, and J. Y. j Hsu, "Sentiment Value Propagation for an Integral Sentiment Dictionary Based on Commonsense Knowledge," in 2011 International Conference on Technologies and Applications of Artificial Intelligence, **2011**, pp. 75–81.

[92] B. Duthil, F. Trousset, G. Dray, J. Montmain, and P. Poncelet, "Opinion Extraction Applied to Criteria," in Database and Expert Systems Applications, **2012**, pp. 489–496.

[93] G. Gezici, R. Dehkharghani, B. A. Yanikoglu, D. Tapucu, and Y. Saygin, “SU-Sentilab: A Classification System for Sentiment Analysis in Twitter.,” presented at the SemEval@ NAACL-HLT, **2013**, pp. 471–477.

[94] J. M. Chenlo and D. E. Losada, “An empirical study of sentence features for subjectivity and polarity classification,” Inf. Sci., vol. 280, no. Supplement C, pp. 275–288, Oct. **2014**.

[95] J. K.-C. Chung, C.-E. Wu, and R. T.-H. Tsai, “Improve Polarity Detection of Online Reviews with Bag-of-Sentimental-Concepts.” Proceedings of the 11th ESWC. Semantic Web Evaluation Challenge, Crete. Springer, pp.379-420, **2014**.

[96] F. Bravo-Marquez, M. Mendoza, and B. Poblete, “Meta-level sentiment models for big social data analysis,” Knowl.-Based Syst., vol. 69, no. Supplement C, pp. 86–99, Oct. **2014**.

[97] D. Reforgiato Recupero, V. Presutti, S. Consoli, A. Gangemi, and A. G. Nuzzolese, “Sentilo: Frame-Based Sentiment Analysis,” Cogn. Comput., vol. 7, no. 2, pp. 211–225, Apr. **2015**.

[98] F. Bisio, C. Meda, P. Gastaldo, R. Zunino, and E. Cambria, “Sentiment-Oriented Information Retrieval: Affective Analysis of Documents Based on the SenticNet Framework,” in Sentiment Analysis and Ontology Engineering, Springer, Cham, **2016**, pp. 175–197.

[99] E. Cambria and A. Hussain, Sentic computing: Techniques, tools, and applications, vol. 2. Springer Science & Business Media, **2012**.

[100] G. Salton and M. Mcgill, Introduction to Modern Information Retrieval. McGraw-Hill, Inc., **1986**.

[101] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, “Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1,” D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, **1986**, pp. 77–109.

[102] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A Neural Probabilistic Language Model,” J. Mach. Learn. Res., vol. 3, no. Feb, pp. 1137–1155, **2003**.

[103] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (Almost) from Scratch,” J. Mach. Learn. Res., vol. 12, no. Aug, pp. 2493–2537, **2011**.

[104] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., **2013**, pp. 3111–3119.

[105] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, “Improving Word Representations via Global Context and Multiple Word Prototypes,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, Stroudsburg, PA, USA, **2012**, pp. 873–882.

[106] T. Buckwalter, “Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02,” ISBN 1-58563-324-0, **2004**.

[107] L. L. Dhande and G. K. Patnaik, “Analyzing sentiment of movie review data using Naive Bayes neural classifier,” Int. J. Emerg. Trends Technol. Comput. Sci. IJETTCS, vol. 3, no. 4, **2014**.

[108] A. Sharma and S. Dey, “An Artificial Neural Network Based Approach for Sentiment Analysis of Opinionated Text,” in Proceedings of the 2012 ACM Research in Applied Computation Symposium, New York, NY, USA, **2012**, pp. 37–42.

[109] S. Jiang, G. Pang, M. Wu, and L. Kuang, “An improved K-nearest-neighbor algorithm for text categorization,” Expert Syst. Appl., vol. 39, no. 1, pp. 1503–1509, Jan. **2012**.

[110] B. Li, “Learning dimensional sentiment of traditional Chinese words with word embedding and support vector regression,” in 2016 International Conference on Asian Language Processing (IALP), **2016**, pp. 324–327.

[111] S. Ginosar and A. Steinitz, “Sentiment Analysis using Linear Regression,” **2012**.

[112] P. Rajan and S. P Victor, “Web Sentiment Analysis: Comparison of Predicted Results with Neural Networks,” vol. 4, May **2014**.

[113] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum, “Introducing the Arabic wordnet project,” presented at the Proceedings of the third international WordNet conference, pp. 295–300, **2006**.

[114] “The Stanford Natural Language Processing Group.” [Online]. Available: https://nlp.stanford.edu/projects/arabic.shtml. **[Accessed: 08-Nov-2017]**.

[115] P. Rosso, Y. Benajiba, and A. Lyhyaoui, “Towards a Measure for Arabic Corpora Quality,” presented at the Proc. 4th Conf. on Scientific Research Outlook & Technology Development in the Arab world, SROIV, Damascus, Syria. pp. 11-14, **2006**.

[116] J.Davis & M. Goadrich, “The relationship between Precision-Recall and ROC curves,” In Proceedings of the 23rd international conference on Machine learning ACM, **2006,** pp. 233-240.

[117] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature,” Geosci. Model Dev., vol. 7, no. 3, pp. 1247–1250, **2014**.

[118] K. Jędrzejewski and M. Zamorski, “Performance of k-nearest neighbors algorithm in opinion classification,” Found. Comput. Decis. Sci., vol. 38, no. 2, pp. 97–110, **2013**.

[119] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer, “Short-term load forecasting in a non-residential building contrasting models and attributes,” Energy Build., vol. 92, no. Supplement C, pp. 322–330, Apr. **2015**.

[120] J.-S. Chou and C.-F. Tsai, “Concrete compressive strength analysis using a combined classification and regression technique,” Autom. Constr., vol. 24, pp. 52–60, Jul. **2012**.

[121] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, “Deep learning for stock prediction using numerical and textual information,” in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), **2016**, pp. 1–6.

[122] A. F. Sheta, S. E. M. Ahmed, and H. Faris, “A comparison between regression, artificial neural networks and support vector machines for predicting stock market index,” Soft Comput., vol. 7, p. 8, **2015**.

[123] M. Kanevski, R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, V. Demyanov, and S. Canu, “Environmental data mining and modeling based on machine learning

algorithms and geostatistics," Environ. Model. Softw., vol. 19, no. 9, pp. 845–855, Sep. **2004**.

[124] G. Li, X. Xing, W. Welsh, and H. Rabitz, "High dimensional model representation constructed by support vector regression. I. Independent variables with known probability distributions," J. Math. Chem., vol. 55, no. 1, pp. 278–303, Jan. **2017**.

[125] J. Gaudart, B. Giusiano, and L. Huiart, "Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data," Comput. Stat. Data Anal., vol. 44, no. 4, pp. 547–570, Jan. **2004**.

[126] A. Koutras, A. Panagopoulos, and I. A. Nikas, "Evaluating the Performance of Linear and Nonlinear Models in Forecasting Tourist Occupancy in the Region of Western Greece," in Tourism and Culture in the Age of Innovation, Springer, Cham, **2016**, pp. 377–391.

[127] F. Sağlam, H. Sever, and B. Genç, "Developing Turkish sentiment lexicon for sentiment analysis using online news media," in 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), **2016**, pp. 1–5.

APPENDEX 1: A SAMPLE FROM OUR GLASC CORPUS FİLES İN EACH CATEGORY

Negative Category
اعتقلت شرطة الاحتلال الاسرائيلية اليوم "الأحد" أربعة شباب فلسطينيين بمدينة القدس للاشتباه في قيامهم بالإخلال بالنظام والاشتباك مع الشرطة الإسرائيلية في الحرم القدسي الشريف في وقت سابق اليوم. ونقلت صحيفة "جيروزاليم بوست" عن الشرطة الإسرائيلية إنه ما أن فتح الموقع المقدس في القدس للزوار اليوم تجمعت مجموعات من الشباب المسلمين وهتفوا بشعارات ضد الزوار اليهود في محاولة لعرقلة عمليات الدخول الروتينية في المجمع. وأشارت الشرطة إلى أنه تم نشر عدد من الضباط في القدس، وخاصة في البلدة القديمة ومناطق الاحتكاك من أجل الحفاظ على النظام وحماية حرية العبادة والأمن في القدس.
صحافة وطنية أحدث شخص يشغل منصب رئيس مصلحة في مديرية المياه والغابات بمدينة تازة فوضى عارمة مساء أمس السبت وذلك بعد ارتكابه لمخالفة سير على الطريق العام بسيارة الخدمة في وقت متأخر من الليل وهو في حالة سكر طافح. وارتكب ذات المسؤول مخالفة بعد تجاوزه لعلامة قف ورفضه الامتثال لشرطيان حاولا تطبيق القانون في حقه حيث امتنع عن تقديم أوراق السيارة لهما وتسجيل المخالفة. وحاول الشرطيان إغلاق الطريق أمامه بركن سيارة أمام سيارته لكن ذات الشخص الذي كان في حالة متقدمة من السكر، قام يصدم السيارة التي وضعت أمامه بطريقة جنونية ونفى توفره على أوراق السيارة
أقدم شاب في العقد الثاني من العمر، على الانتحار بربط قطعة قماش حول عنقه، صباح اليوم الأحد، شرق محافظة القنفذة. وقال الناطق الإعلامي بشرطة منطقة مكة المكرمة العقيد دكتور عاطي بن عطيه القرشي لـ"المواطن"، إنه في صباح يوم الأحد 19 شوال 1437، أُبلِغت الجهات الأمنية بشرطة محافظة القنفذة، عن وجود شخص متوفى بمنزله. وأضاف أنه انتقل رجال الضبط الجنائي والجهات الأخرى ذات العلاقة إلى الموقع، وبعد إجراءات المُعاينة والفحوصات الأولية، اتضح أن المتوفى في العشرين من العمر، قام بإيذاء نفسه، وذلك بربط شماغ حول عنقه، مما أدى إلى وفاته، وتم حفظ الجثمان بالمستشفى، وإحالة ملف القضية لجهة الاختصاص.
لم يقع حل أي مؤسسة استعلامتية أكد وزير الداخلية الأسبق والنائب بكتلة حركة النهضة، على العريض، أنه لم يقع حل أي مؤسسة من المؤسسات الإستعلاماتية. وقال في مداخلة في جلسة برلمانية عامة حول ملف اغتيال الشهيدين شكري بلعيد ومحمد البراهمي أن هذه المؤسسات تم تدعيمها ولم يقع حلها عام 2012، عندما كان على رأس وزارة الداخلية. وأضاف أن منظومة "البوليس السياسي" تم حلها في مارس 2011 إبّان فترة "فرحات الراجحي". وفي سياق آخر، لاحظ على العريض أن هناك تهديدات تستهدفه من مجموعات عديدة مثل تنظيم أنصار الشريعة، منذ أن كان وزيرا. وأقر على العريض أن الاغتيالات التي قامت بها المجموعات الإرهابية هي اغتيالات سياسية.
فشلت مؤشرات البورصة في الحفاظ على مكاسبها الصباحية وتحولت إلى الخسائر مع نهاية جلسة تعاملات اليوم في مستهل التعاملات الأسبوعية متأثرة بمبيعات المصريين. حولت القيمة السوقية للأسهم أرباحها إلى خسائر مسجلة 1.9 مليار جنيه لتغلق عند مستوى 397.209 مليار جنيه. تراجع مؤشر إيجي إكس 30 الذي يقيس أداء أنشط ثلاثين شركة بنحو 82 نقطة بنسبة 1.1%، ووصل إلى مستوى 7337 نقطة. هبط مؤشر إيجي إكس 70 الذي يقيس أداء الأسهم الصغيرة والمتوسطة بنسبة 0.72 %، وأغلق مؤشر إيجي إكس 100 الأوسع نطاقا والذي يضم الشركات المكونة لمؤشري ايجي اكس 30 وايجي اكس 70 على تراجع بنسبة 0.72%. فشلت مشتريات الأجانب والعرب في تماسك السوق وحققوا صافي عمليات شرائية بلغت 17 مليون جنيه، مقابل عمليات بيعة للمستثمرين المصريين، وسط قيمة تداولات متدنية بلغت 270 مليون جنيه. سيطر اللون الأخضر على أسهم 32 شركة، من إجمالي 179 شركة تم التداول عليهم خلال الجلسة، فيما تراجعت أسهم 102 شركة.
حوادث 25 مليار جنيه خسائر البورصة خلال 3 شهور منيت البورصة المصرية خلال الربع الثاني من العام الحالي(اول ابريل-نهاية يونيو) بخسائر بلغت نحو 25 مليار جنيه ليبلغ رأسمال السوقي لأسهم الشركات المقيدة بالبورصة نحو 382.5 مليار جنيه مقارنة 407.5 مليار جنيه خلال الربع السابق له بهبوط بلغ نحو 6.1 %. وأظهر التقرير الربع السنوي للبورصة المصرية تراجع مؤشرات السوق الرئيسية والثانوية حيث هبط مؤشر السوق الرئيسي "إيجي اكس 30" بنحو 7.74 % ليبلغ مستوى 6943 نقطة, كما تراجع مؤشر الأسهم الصغيرة والمتوسطة "إيجي اكس 70" بنحو 4.42 % ليغلق عند مستوى 351 نقطة , شملت التراجعات مؤشر "إيجي اكس 100" الأوسع نطاقا والذي فقد نحو 6.12 % ليغلق عند مستوى 744 نقطة. وأشار التقرير إلى تراجع إجمالي التداولات خلال 3 شهور لتبلغ 60.6 مليار جنيه, في حين بلغت كمية التداول 12.426 مليون ورقة منفذة على 1.261 مليون عملية مقارنة بنحو 67.9 مليار جنيه وكمية تداول بلغت 17.813 مليون جنيه منفذة على 1.592 مليون عملية خلال الثلاثة شهور السابقة عليها. وفيما يتعلق ببورصة النيل فقد سجلت قيمة تداول بلغت 250.6 مليون جنيه وكمية تداول بلغت 138.4 مليون جنيه ورقة منفذة على 30.916 ألف عملية خلال الثلاث شهور, وقد استحوذت الأسهم على 61.12 % من إجمالي التداول داخل المقصورة ,في حين استحوذت السندات على نحو 38.88 % خلال 3 شهور. وأضاف التقرير أن تعاملات المستثمرين المصريين استحوذت على 82.59 % من إجمالي تعاملات السوق فيما استحوذ المستثمرون الأجانب غير العرب على 11.76 %, والمستثمرون العرب على 5.65 % وذلك بعد استبعاد الصفقات. وقد سجل

الأجانب غير العرب صافي شراء بلغ 1.801 مليار جنيه خلال الثلاثة شهور، بينما سجل العرب صافي شراء بقيمة 335.81 مليون جنيه وذلك بعد استبعاد الصفقات. يذكر أن صافي تعاملات الأجانب غير العرب سجلت صافي شراء بلغ 1.399 مليار جنيه منذ بداية العام بينما سجل العرب صافي شراء قدره 930.48 مليون جنيه خلال نفس الفترة وذلك بعد استبعاد الصفقات. ولفت إلى أن المؤسسات استحوذت على 60.93 % من المعاملات في البورصة، وكانت باقي المعاملات من نصيب الأفراد بنسبة 37.07 % . وسجلت المؤسسات صافي شراء بقيمة 622.84 مليون جنيه خلال 3 شهور وذلك بعد استبعاد الصفقات. وفي سوق السندات بين التقرير أن إجمالي التداولات عليها بلغ 22.442 مليار جنيه خلال 3 شهور كما بلغ إجمالي التعامل على السندات نحو 21.945 مليون سند.
Positive Category
مكررات الأرباح الأسعار المفكرة النتائج المالية القوائم المالية قائمة كبار الملاك السوق السعودي: 12 شركة أعلنت عن نتائجها المالية للنصف الأول 2016 خلال يومي الأحد والإثنين 18-07-2016 أرقام - خاص أعلنت خلال يومي الأحد والإثنين، 12 شركة مدرجة في السوق السعودي عن نتائجها المالية للنصف الأول 2016، كان ابرزها البنك الأهلي ومصرف الراجحي ومعادن. وقد ارتفعت أرباح 8 شركات من بين تلك الشركات، بينما تراجعت أرباح 3 شركات منها، وتكبدت شركة واحدة (وهي سند للتأمين) خسائر فصلية، كما يتضح أدناه: الشركات المعلنة عن نتائجها المالية (مليون ريال) الشركة 6 أشهر-2015
أظهر استطلاع للرأي، أجرته مجلة "فرانس فوتبول" الفرنسية، أن أغلبية الفرنسيين يرشحون المنتخب الألماني، للفوز ببطولة كأس الأمم الأوروبية، المقامة في فرنسا حتى 10 يوليو/تموز المقبل. ووصل المنتخب الألماني، إلى ربع نهائي البطولة، في مواجهة المنتخب الإيطالي، وفي حالة تخطيه عقبة الأزوري، فإنه سيواجه في نصف النهائي على الأرجح، المنتخب الفرنسي، صاحب الترشيحات الأكبر للفوز على أيسلندا في ربع النهائي. وسألت المجلة الفرنسية، قراءها: "هل المنتخب الألماني هو المرشح الأول للفوز بلقب بطولة أمم أوروبا لكرة القدم؟"، مشيرة إلى أن منتخب ألمانيا، يمتلك كل المقومات الأساسية للفوز باللقب الرابع في تاريخ المانشافت وتحقيق رقم قياسي جديد في الفوز بكأس أوروبا. كما أوضحت أن المانشافت لديه الخبرة اللازمة التي تؤهله للفوز باللقب، نظرًا لكونه بطل كأس العالم الأخيرة، ويمتلك أفضل حارس مرمى في العالم، مانويل نوير، وأفضل اللاعبين في كل المراكز. وكشفت المؤشرات الأولية لنتائج الاستطلاع الذي شارك فيه 2516 شخصًا، عن موافقة 58% من مجموع الأصوات، على كون ألمانيا المرشح الاول للفوز باللقب، بينما رفض 42% من المصوتين ذلك، مؤكدين أن منتخبا آخر غير ألمانيا سيتوج باللقب.
دُبي ترنيم محمد تعتبر دبي واحدة من أجمل دول العالم العربية على الإطلاق، حيث إنها غنية بالمعالم السياحية ومراكز التجارة الاقتصادية، التي تجعلها من أهم عوامل جذب انتباه السياح إليها من كل مكان بالعالم. فإن لم تزر دُبي من قبل، فالآن نُقدم لك 4 أمور تشجعك على زيارة هذه الإمارة الرائعة على الفور: زيارة مدن الملاهي: تمتلىء دُبي بالعديد من مدن الملاهي غير التقليدية؛ سواء مدينة الملاهي المائية أو الثلجية، والتي تتيح لزائر دُبي التمتع بالثلج في ظل الطقس الحار الذي تعاني منه الإمارة. رحلات السفاري: اقتحام صحراء الخليج العربي في جولة من جولات السفاري، من أكثر الأمور التي يُمكنك التمتع بها في دُبي، فمن خلال سيارات الدفع الرباعي بإمكانك القيام بمغامرة رائعة وسط الرمال، بالإضافة إلى التخييم في الصحراء، والتمتع بالهدوء والبعد عن ضوضاء المجتمعات العمرانية. حسن الاستقبال والضيافة: تُخضع إمارة دُبي الرائعة جميع إمكانيتها من أجل راحة زوارها، ولتترك في أذهانهم انطباعًا جيدًا يدوم إلى الأبد، لذا فلن تجد حفاوة استقبال وترحيب كتلك التي تجدها في دُبي. الطقس الرائع: تتمتع إمارة دبي بالطقس المعتدل الدافىء شتاءً، والذي سينسيك الطقس البارد في بلدك، كما تطل حدود دُبي على مياه شواطىء الخليج العربي الهادئة النظيفة.
فازت شركة الإمارات العالمية للألومنيوم، أكبر منتج للألومنيوم الأولي في منطقة الخليج، بجائزة الدرع الذهبية للتميز في الاستدامة البيئية من المنظمة العربية للمسؤولية الاجتماعية، تقديراً لجهودها الدؤوبة والمتواصلة بمجال حماية البيئة. وتهدف جوائز المنظمة العربية للمسؤولية الاجتماعية إلى تكريم الأداء الاستثنائي بمجال المسؤولية الاجتماعية، حيث تسلط الضوء بشكل خاص على المبادرات البيئية البارزة التي يتم تدشينها ودعمها من جانب الشركات العامة والخاصة في الدول الأعضاء بجامعة الدول العربية. وتشمل هذه المبادرات قطاعات الابتكار التكنولوجي، والمباني الخضراء، والحفاظ على الطاقة، والتنمية البيئية والاستدامة، وجهود التنظيف البحرية، والاقتصاد الأخضر. وقال عبدالله كلبان، العضو المنتدب والرئيس التنفيذي: «تلتزم الإمارات العالمية للألومنيوم بالعمل على تقليص آثار عملياتنا على البيئة، وخلال سعينا لتحقيق التحسين المستمر في كل ما نقوم به، نحرص على تطبيق وتبني أفضل الممارسات في مجال الإدارة البيئية، حتى نكون نموذجاً يحتذى به بمجال التنمية المستدامة.
اخبار العراق الان 0 صرّح رئيس لجنة الاستثمار في مجلس محافظة كربلاء الاثنين، ان هيئة الاستثمار الوطنية متمثلة بالسيد سامي الاعرجي، قد واعدت المحافظة بإنجاز إجازة الاستثمار الخاصة بمطار كربلاء (الامام الحسين) الدولي، نهاية شهر آب. وقال زهير ابو دكه في حديث لوكالة نون الخبرية، ان "هيئة الاستثمار الوطنية قبل عام ونصف من الان، ان تمنح اجازة الاستثمار للبدء بإنجاز مطار #كربلاء الدولي، الا اننا لحد الان لم نرَ أي شيء ملموس على الواقع، الا ان الزيارة الاخيرة للسيد سامي الاعرجي الى مجلس المحافظة قد واعد بأن يتم منح الاجازة نهاية شهر آب الجاري". وأضاف ابو دكه، ان "في حالة عدم تنفيذ الهيئة لوعودها فإن لمحافظة كربلاء ، خطوات اخرى سنعمل عليها ولكل حادث حديث". واتهم وزير النقل السابق المهندس عامر عبد الجبار هيئة الاستثمار الوطنية، في (31 تموز 2016)، بعرقلة انشاء مطار الفرات الاوسط (الامام الحسين)، داعيا رئيس الوزراء حيدر العبادي لوضع حلول مناسبة لإدارة الهيئة لإنجاح المشاريع الاستثمارية المعرقلة في العراق وخاصة التي يتم بيعها بالباطن.
تقييمات المكنسة الذكية «روبوت» من «أل جي» تؤهلها للأعلي مبيعات

صورة ارشيفية منة يحيي   في عدد من المراجعات الخاصة بالتقارير الصحفية حصلت المكنسة الكهربائية روبوت لشركة إل جي على تقييمات متميزة من وسائل الإعلام الأمريكية وهو السوق الأكبر للأجهزة الترفيهية الذكية مما يعني تأهيلها للحصول على الأعلى مبيعات والأفضل. ونشرت وكالة الأنباء الكورية "يونهاب" تقرير، ذكرت فيه أن مجلة فوربس الأمريكية، أشادت بالمكنسة لتضمينها العديد من الميزات بما في ذلك تنويع الوظائف المختلفة للتنظيف حسب نوع الأرضيات مثل الأرضيات الخشبية وأرضيات السجاد، واستشعار لتجنب العقبات. في حين قدمت مجلة Reviewed.com الخاصة بالمراجعة التابعة لصحيفة يو إس إيه توداي الأمريكية، 9.1 درجة من أصل 10 درجات إلى المكنسة الكهربائية الروبوتية RoboKing ، مضيفة أنها تتميز بالتصميم لتحسين تنظيف الزاوية في المنزل فضلا عن قوة الشفط لها. وأضافت أن المكنسة RoboKing لها وظيفة جيدة في تنظيف السجاد خلافا لمنافساتها التي تواجه صعوبات فيه. وهو ما دعي مسئول في شركة إل جي العالمية، يؤكد قيادة الشركة للسوق العالمي للمكنسة الكهربائية الروبوتية من خلال الوظائف المتميزة للتنظيف والأداء الذكي.

Neutral Category
دمشق تؤكد للامم المتحدة انها ستشارك بمحادثات جنيف المرتقبة   نيويورك 31 تموز (بترا)-ابلغت الحكومة السورية الامم المتحدة اليوم الاحد، انها ستشارك في محادثات جنيف المرتقبة في آب المقبل كمحاولة جديدة لايجاد حل سياسي للنزاع السوري المستمر منذ اكثر من خمسة اعوام. وقال نائب المبعوث الخاص للامم المتحدة الى سوريا، رمزي عز الدين رمزي، عقب لقائه وزير الخارجية السوري وليد المعلم ونائبه فيصل المقداد، "اكد لي الوزير ان الحكومة السورية على موقفها من انها ستشارك في المحادثات المنتظر عقدها في خلال اسابيع بنهاية آب المقبل". وكان دي ميستورا قد اعرب الثلاثاء الماضي، عن امله باستئناف محادثات السلام السورية "اواخر اب"، في ختام اجتماع في جنيف مع مسؤولين اميركيين وروس.
أشارت صحيفة ليكيب الفرنسية اليوم الخميس، إلى أن نجم خط وسط باريس سان جيرمان الفرنسي، الإيطالي ماركو فيراتي، سيرث الرقم 10 من نجم الفريق المنتقل حديثاً إلى صفوف مانشستر يونايتد الإنجليزي، السويدي زلاتان إبراهيموفيتش. ونافس فيراتي على الرقم زميله الأرجنتيني في الفريق خافيير باستوري، الذي تربطه شائعات في إيطاليا بالعودة للعب في "الكالتشيو"، وبالتحديد في صفوف نادي ميلان. وبذلك أنهى فيراتي كل الأخبار التي أكدت في الأسابيع القليلة الماضية، أن اللاعب قد يغادر سان جيرمان في الانتقالات الصيفية الجارية، بسبب الإغراءات التي تصله من إنجلترا وإسبانيا.
الفنانة المصرية تؤكد أن موقع الصور "انستجرام" هو المفضل لديها وتنشر من خلاله أخبارها وصورها. المصدر: القاهرة- من دعاء السيد نفت الفنانة المصرية، مي عز الدين، امتلاكها صفحة على موقع التواصل الاجتماعي “فيس بوك”، بعد نشر أخبار لها بوسائل الإعلام منسوبة لصفحات تحمل اسمها. وقالت مي إن موقع الصور “انستجرام” هو المفضل لديها، والذي تنشر من خلاله أحدث أخبارها وصورها. وأضافت النجمة المصرية الشابة عبر حسابها في موقع “انستجرام” إنها تمتلك حساباً على “تويتر”، ولكنها نادراً ما تستخدمه, مشددةً أنها غير مسؤولة عن الصفحات التي تحمل اسمها. ومن جانب آخر، تقرأ مي عز الدين حالياً مجموعة من السيناريوهات، لتختار أفضلهم، بعد النجاح الذي حققه مسلسلها “دلع بنات” خلال رمضان الماضي
توشاك للاعبي الوداد: الرسمية مسألة اجتهاد وليست احتكارا لأي كان تعليق حسم جون توشاك، مدرب الوداد الرياضي، في مسألة اختياراته التقنية، من خلال رسالة واضحة للاعبيه تدعوهم إلى البذل والعطاء بغية نيل الرسمية والحفاظ عليها في المباريات المقبلة. وقال توشاك للاعبيه، على هامش الحصة التدريبية، لأمس السبت، وفق ما نشره الموقع الرسمي للفريق، إن المشاركة يجب أن تكون عن جدارة واستحقاق، من خلال بذل مجهودات كبيرة قصد نيل الرسمية، والمشاركة في المباريات. وتابع “مباراتنا أمام زيسكو أهم من سابقتها ضد الأسيك، لأنها بمثابة تأكيد للأولى، وضياع الثلاث نقط فيها سيجعل الفوز، الذي حققناه في ساحل العاج من دون أهمية، وأؤكد لكم مرة أخرى أن الرسمية مسألة اجتهاد، وليست احتكارا لأي كان”.
"الزراعة" تتسلم 20 الف طن قمح وشعير .عمان 23 اب(بترا)-. تسلمت اللجنة المركزية لشراء الحبوب المحلية في وزارة الزراعة من المزارعين 20179 طنا من القمح والشعير منها 11429 طن قمح مواني و8750 طن شعير علفي عقب اعلان اللجنة انتهاء اعمال استلام المحاصيل اليوم. وقالت اللجنة خلال اجتماعها، اليوم الثلاثاء، برئاسة أمين عام الوزارة الدكتور راضي الطراونة ان اللجنة تسلمت من المزارعين في مراكز الاستلام (الشمال، الوسط، الجنوب) 804 طن بذار قمح و(690) طن بذار شعير من خلال لجان استلام البذار في الأقاليم الثلاثة. واضافت ان الكلفة المالية الإجمالية للكميات المستلمة 7ر600 مليون دينار، وبدات اللجان بصرف المستحقات المالية للمزارعين على ان تنتهي خلال أسبوعين من تاريخ انتهاء عمليات الاستلام.
أطلقت شركة مايكروسوفت أمس إطلاقها خدمة مايكروسوفت ستريم Microsoft Stream، وهي خدمة جديدة تسمح باستعمالها لتحميل ومشاركة ملفات الفيديو مع زملاء العمل. وعملت الشركة على إتاحة الخدمة الجديدة للمعاينة والاستخدام ابتداء من أمس، وتتماشى الخدمة الجديدة مع خدمة الفيديو الموجودة حاليا Office 365 Video، سواء كانت الشركات تستخدم خدمة Office 365 أم لا. وأشار نائب رئيس الشركة لمجموعة منتجات الذكاء من مايكروسوفت جيمس فيليبس في تدوينة نشرها إلى أنه يمكن لأي مستخدم يمتلك عنوانا بريديا الكترونيا تجاريا تسجيل الدخول ومعاينة الخدمة واستعمالها من أجل تحميل ومشاركة ملفات الفيديو مع زملاء العمل في مكان العمل. وتستغرق عملية الاشتراك بضع ثوان، ويمكن تحميل مقاطع الفيديو بكل سهولة عن طريق السحب والإسقاط.

APPENDEX 2: A SAMPLE FROM OUR AR-SENTİCNET CONCEPT BASED SENTİMENT ANALYSİS LEXİCON FOR ARABİC

Senses set ; English concept ; Sentiment score; Arabic concept	No
آلية ; 0.053 ; mechanism ; جهاز ميكانيكي ; آلة معقدة ; اختراع الإنسان ; آلة ; جهاز ; ميكانيكية ; آليات	1
آلية ضبط الوقت ; -0.57 ; timekeeping mechanism ; راقب ; ساعة التوقيف ; ساعة ; ساعة اليد	2
بالغ ; 0.439 ; adult ; كائن بشري ; رجل ; النساء ; الذكر ; جنسي، ومان ; راشد ; فرط	3
بالمعنى الفطري ; 0.872 ; innate sense ; الصراحة ; استقامة ; صفة مميزة ; الفورية ; مباشرة	4
تأجير ; 0.038 ; rent out ; منزل ; سكن ; بناء متعدد الطوابق ; عقد الإيجار ; منزل كبير ; ايجار ; مدة الإيجار ; فترة الإيجار ; إيجار	5
تأخذ وقتا ; 0.586 ; take up time ; متأخر ; دين ; الموعد النهائي ; ضغط عصبى ; القليل من الوقت	6
ثابت ; -0.04 ; stationary ; مستقر ; غير متحرك ; تسوية ; استقرار ; أكيد ; محقق ; مؤكد	7
ثبات الغرض ; 0.675 ; firmness of purpose ; قوة الإرادة ; سوف السلطة ; حسم ; سمة ; ثبات	8
جائزة ; 0.883 ; award ; ميزة ; تفوق ; أهمية ; علاوة ; مكافأة	9
جدير بالملاحظة ; 0.869 ; noticeable ; قابل للتمييز ; ملموس ; واضح ; كشف ; يمكن إدراكه	10
حرر ; 0.913 ; liberate ; حر ; إطلاق سراح ; زحزح ; خلص ; أزاح ; أزال القيود	11
حركة الهواء; -0.02 ; air movement ; معتدل ; نسيم ; عاصفة إستوائية ; ينفخ ; عاصف	12
خجل ; -0.54 ; abash ; غير ملائم ; لا يحسد عليها ; إلى حد مربك ; إحراج ; خجول ; احمر	13
خدمة سيئة ; -0.65 ; bad service ; الموظفين سيئة ; فندق رخيص ; مطعم رخيص ; الموظفين غير ودية ; وقت الانتظار الطويل	14
دائم ; -0.79 ; lasting ; أبدي ; مكانة ; خالد ; مستمر	15
دخان السجائر ; -0.03 ; cigarette smoke ; دخان ; التبغ ; الدخان التبغ ; سرطان الرئة ; سيجار	16
ذكي ; 0.921 ; intelligent ; نباهة ; فضولي ; تصبح حكيم ; فهم	17
ذو قيمة ; 0.143 ; valuable ; الدولار ; دقة ; غنى ; تغير بسيط ; ثروة	18
رأي ; 0.142 ; opinion ; السبب ; التفكير المسبق ; يعتبر ; احساس قوي ; خيار ; رأي قانوني ; حكم ; فكرة ; نصيحة ; نصح ; إقتراح	19
رؤية ايجابية ; 0.776 ; positive outlook ; الرضا الذاتي ; استمتع بالحياة ; شعور جيد ; سلوك جيد ; استمتع	20
زائد ; 0.664 ; plus ; مفيد ; إيجابي ; جيد ; تمييزي ; متزايد	21
زميل متآمر ; -0.54 ; fellow conspirator ; جريمة ; قتل ; متواطئ ; شريك في الجريمة	22
سؤال ; 0.609 ; question ; إجابة ; تحقيق ; الإجابة على السؤال ; الرد ; طلب ; استجواب ; تساؤل ; استعلام ; استنطاق ; مسألة	23
ساعة التوقيف ; 0.532 ; stopwatch ; ساعة حائط ; الساعة الرملية ; راقب ; ساعة ساعة اليد	24
شاب ; 0.025 ; young man ; شخص ; على قيد الحياة ; مستثمر ; كائن بشري ; بستاني ; فتى ; ناشئ ; حدث ; يافع ; صبي ; ولد	25
شجرة تفاح ; 0.059 ; apple tree ; نبات ; شجرة ; فاكهة ; تفاحة ; أحمر	26
صانع ; 0.089 ; maker ; بناء ; عامل بناء ; فلاح ; الموظف ; باني ; مصنع ; منتج ; مؤسسة انتاجية ; حرفي ; صنائعي ; باري ; مخترع ; خالق	27
صالون الشعر ; 0.078 ; hair salon ; الحلاق ; صالون تصفيف الشعر ; صالون تجميل ; صالون	28
ضار ; -0.40 ; maliciously ; خبث ; حقد ; الإيذائية ; مؤذ ; مسيئ ; مجحف ; مضر	29
ضباب الصباح ; -0.03 ; morning mist ; ضبابي ; شيء ضبابي ; غائم ; غير واضحة ; ضبابية	30
طامع ; -0.43 ; covetous ; الغيرة ; الحسد ; حسود ; بطمع ; حاسد	31
طعام لذيذ ; 0.634 ; delicious food ; لذيذ ; وجبة ساخنة ; طعام جيد ; المنتجات الطازجة	32
ظرف ; 0.052 ; envelop ; فتح الرسالة ; بريد ; طابع بريدي ; نشرة ; غلاف ; مغلف	33

34	ظلام في الخارج ; -0.04 ; dark outside ; يظلم ; وقت الليل ; نهاية اليوم ; غروب ; الغسق
35	عائق ; -0.68 ; draw back ; تراجع ; متراجع ; دعم ; نقل ; استسلم ; عقبة ; مانع ; حاجز ; عسر ; مشقة ; صعوبة ; عائقة ; عرقلة
36	عادة صحية ; 0.831 ; healthy habit ; والحفاظ على صحة الجلد ; ابقى بصحة جيدة ; حافظ على لياقتك ; والحفاظ على صحة جيدة ; الحفاظ على قوة العضلات
37	غامض ; -0.72 ; inscrutable ; غير قابل للتفسير ; غير المبررة ; محير
38	غسيل ملابس ; 0.549 ; laundry ; الملابس النظيفة ; غسيل الملابس ; مسحوق تنظيف ; شماعات الملابس ; شراء مسحوق الغسيل
39	فارغة ; -0.02 ; empty ; زجاج فارغ ; مساحة فارغة ; عديم القيمة ; معدة فارغة ; كون ; فارغ ; خاو
40	فرح عظيم ; 0.132 ; great joy ; الشعور بالفخر ; سعادة ; فرحة الشعور ; إستمتع ; يشعر متعة
41	قابل للكسر ; -0.06 ; breakable ; هش ; ضعيف ; من السهل كسر ; الواهية
42	قاحل ; -0.79 ; barren ; الصحراء ; مجدب ; رمل ; أرض قاحلة ; صحراء ; جاف
43	كبر ; 0.287 ; grow old ; نباهة ; تصرف بنضج ; أخبار جيدة ; تعلم بسرعة ; التي لا نهاية لها ; نبت ; نمى ; نضج ; وسع ; تقدم العمر ; سن ; شيخوخة ; هرم
44	كسب المال ; 0.781 ; money earn ; الراتب الوظيفي ; اموال اضافية ; عائد الاستثمار ; مكاسب مالية ; تحقيق مكاسب مالية
45	لا داعي ; -0.47 ; needless ; تافه ; غير ضروري ; خامل ; بدون فائدة ; بلا هدف
46	لامع ; -0.51 ; glossy ; ساطع ; براق ; ومضة ; متلألئ
47	مأساوي ; -0.57 ; tragic ; الإلياذة ; مأساويا ; الممثلة التراجيدية ; مأساة ; حزن
48	مثير للدهشة ; 0.357 ; amazingly ; مدهش ; مندهش ; مفاجئ ; بشكل مفاجئ
49	ناجح ; 0.903 ; successful ; مزدهر ; بهي
50	نزع السلاح ; -0.82 ; disarm ; ثكل ; مصادرة ; تحرم ; جرده من ملابسه ; يأخذ ; جرد من السلاح ; أخضع للإدارة المدنية ; تجريد من السلاح
51	هادئ ; 0.54 ; quiet ; هدوء ; راحة نفسية ; سكن ; الهدوء
52	هدية عيد ميلاد ; 0.904 ; birthday gift ; هدية ; يوم الاجازة ; هدية مجانية ; مفاجأة ; عيد الميلاد
53	وئام ; 0.819 ; harmoniously ; صحيح ; نقي ; متناسق ; المنسقة
54	وثيقة قانونية ; 0.072 ; legal document ; دعوى ; الإجراءات القضائية ; تأمين ; عقد التغيير ; أمين صندوق ; مستند رسمي ; مستند قانوني ; سند ; صك ; وثيقة رسمية
55	يأذن ; 0.588 ; authorize ; يسمح ; يعطى ; مناسب ; البشري ; ثقة ; أذن ; سمح ; أجاز ; وافق ; صرح ; رخص ; قبل ; أيد ; شجع
56	يحرز هدف ; 0.239 ; score goal ; أحرز هدفاً ; جعل نقطة ; تحقيق الهدف ; نجاح ; نقطة النتيجة

CURRICULUM VITAE

Credentials

Name, Surname: Ahmed NASSER

Place of Birth: Baghdad / Iraq

Marital Status: Married

E-mails: ahmed.r.nasser1984@gmail.com

Address: Department of Computer Engineering, Hacettepe- University, Beytepe, Ankara / TURKEY

Education

High School: Al-najah Secondary School, Baghdad / Iraq

BSc.: Control and Computer Engineering Department, University of technology, Baghdad / Iraq

MSc.: Department of Computer Engineering, Istanbul University, Istanbul / TURKEY

PhD.: Department of Computer Engineering, Hacettepe University, Ankara / TURKEY

Foreign Languages

Arabic (Native), Turkish (Advanced), English (Advanced)

Work Experience

[2007 – to 2012]: Engineer at Control and Computer Engineering Department, University of technology, Baghdad / Iraq

[2012– to 2014]: Lecturer at Control and Computer Engineering Department, University of technology, Baghdad / Iraq

Area of Experience

Natural Language Processing, Machine Learning, Sentiment Analysis, Information Retrieval, Data Base Systems, Operating Systems, Electronic and Embedded systems.

Projects and Budgets

-

Publications

- Ahmed Nasser, Kivanç Dinçer, and Hayri Sever. "Investigation of the Feature Selection Problem for Sentiment Analysis in Arabic Language." Research in Computing Science 110 (2016): 41-54.
- Nasser, A., Sever, H. and Raghavan, V.V. (2017). Utilization of Rough Sets for Intrusion Detection, 17th World Congress of IFSA-SCIS'17, Otsu, Japan.
- Ahmed Nasser and Hayri Sever, “A Large-Scale Arabic Sentiment Corpus Construction Using Online News Media”, Journal of Engineering and Applied Sciences, Vol. ??, (2017), No. ?, pp:??-??. (in press)
- Ahmed Nasser and Hayri Sever, “A Concept-based Sentiment Analysis Approach for Arabic”, International Arab Journal of Information Technology, (2017), (Submitted)

Oral and Poster Presentations

-

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
TO THE DEPARTMENT OF COMPUTER ENGINEERING

Date: 04/01/2018

Thesis Title / Topic: **LARGE-SCALE ARABIC SENTIMENT CORPUS AND LEXICON BUILDING FOR CONCEPT-BASED SENTIMENT ANALYSIS SYSTEMS**

According to the originality report obtained by myself/my thesis advisor by using the *Turnitin* plagiarism detection software and by applying the filtering options stated below on 14/12/2017 for the total of 120 pages including the a) Title Page, b) Introduction, c) Main Chapters, d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 8 %.

Filtering options applied:

1. **Bibliograph**y/~~Works Cited~~ excluded
2. **Quotes excluded** / ~~included~~
3. **Match size up to 5 words excluded**

I declare that I have carefully read Hacettepe University Graduate School of Sciene and Engineering Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Date and Signature

Name Surname:	**Ahmed NASSER**
Student No:	**N13148377**
Department:	**Computer Engıneerıng**
Program:	**Computer Engineering-Doctor of Philosophy (Ph.D.)**
Status:	□ Masters ☒ Ph.D. □ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED.

Prof. Dr. Hayri SEVER

(Title, Name Surname, Signature)