A Novel Hybrid Sentiment Analysis Classification Approach for Mobile Applications Arabic Slang Reviews

Rabab Emad Saudy[1]
Department of Information Systems Technology
Faculty of Graduated Studies for Statistical Research
Cairo University, Cairo, Egypt

Alaa El Din M. El-Ghazaly[2]
Department of Computer and Information Sciences
Sadat Academy for Management Sciences, Cairo, Egypt

Eman S. Nasr[3]
Independent Researcher
Cairo, Egypt

Mervat H. Gheith[4]
Department of Computer Science
Faculty of Graduated Studies for Statistical Research
Cairo University, Cairo, Egypt

***Abstract*—Arabic language incurs from the shortage of accessible huge datasets for Sentiment Analysis (SA), Machine Learning (ML), and Deep Learning (DL) applications. In this paper, we present MASR, a simple Mobile Applications Arabic Slang Reviews dataset for SA, ML, and DL applications which comprises of 2469 Egyptian Mobile Apps reviews, and help app developers meet user requirements evolution. Our methodology consists of six phases. We collect mobile apps reviews dataset, then apply preprocessing steps, in addition perform SA tasks. To evaluate MASR datasets, first we apply ML classification techniques: K-Nearest Neighbors (K-NN), Support vector machine (SVM), Logistic Regression (LR), and Random Forest (RF), and DL classification technique: Multi-layer Perceptron Neural Network (MLP-NN). From the examination for pervious classification techniques, we adopted a hybrid classification approach combined from the top two ML classifier accuracy results (LR, RF), and DL classifier (MLP-NN). The findings prove the adequacy of a hybrid supervised classification approach for MASR datasets.**

Keywords*—Arabic sentiment analysis; mobile application; hybrid classification model; hybrid supervised classification approach; Google play store; random forest; logistic regression; neural network; multi-layer perceptron neural network; machine learning; deep learning

I. INTRODUCTION

Mobile app stores supply an amazingly wealthy source of information on app specification, characteristics, and utilize, and analyzing these information supplies knowledge and a more profound comprehension of the idea of apps. However, manual analysis of this tremendous measure of information on mobile apps is anything but a basic and clear task; it is expensive as far as human effort and time [1]. There are different mobile app stores, for example, Google, and Apple app store, and others that include free and paid mobile apps [2].

Mobile app classification phase is classified based on a significant category or class. In case users want to investigate and discover an app reasonable for their requirements, it is more helpful to have a special predefined classification scheme by which all apps are classified [3].

Being a significant provenance of data for organizations, the requirement to produce exact SA is a significant issue. Most sentiments accumulated from Arabic resources like social media is in colloquial Arabic, as the utilization of Modern Standard Arabic (MSA) in online is uncommon [4].

A few researches have been directed to analyze English mobile apps [5] [6] [7] [8] [9] [10]. In addition, according to the literature review, few researches have analyzed Islamic Arabic mobile apps and Saudi governmental services mobile apps [1] [11] [12]. However, no previous study has constructed, classified or analyzed Egyptian Dialect Arabic (DA) mobile apps reviews dataset.

The contributions in this research can be summed up as follows:

1) Introduce present MASR, simple Mobile Applications Arabic Slang Reviews of Egyptian reviews dataset for SA, ML and DL applications.

2) Investigate the structure, properties of the dataset, and perform tests on selected attributes for sentiment polarity classification.

3) Apply a various supervised ML, DL classifiers to the simple MASR that we gathered.

4) Adopted a hybrid supervised sentiment analysis classification approach including heterogenous approaches: Machine Learning (ML) approach such as: Logistic Regression (LR), and Random Forest (RF), and Deep Learning (DL) approach: Multi-layer Perceptron Neural Network (MLP-NN) classifiers to enhance the performance models of predicting MASR datasets and accuracy.

5) Compare our proposed model approach performance with various ML, and DL models.

The rest of the paper is organized as follows: Section II presents the literature review. Section III presents the six

phases of our proposed hybrid classification approach methodology. Section IV presents experimental results and discussion. Section V presents conclusion and Section VI presents future works.

II. Literature Review

Slight endeavors have been made to anatomize mobile apps reviews to handle mobile apps requirements evolution, advancement information and significant software. Related previous studies handle many aspects in mining mobile apps reviews for different sentiment analysis purposes such as building lexicons, classifying non-functional requirements, classify buggy apps, recognizing high-rated apps, and hybrid system to find the most similar word in lexicon for Egyptian Arabic tweets.

1) Arabic sentiment analysis tasks: El-Beltagy et al. [13] build a sentimental Egyptian Dialect lexicon. Their tests showed that their proposed methodology gave improved results with regards to twitter even with the poor utilized resources.

Fu et al. [14] dealt with an enormous user reviews dataset including about 13 million mobile apps reviews from google play store. The creators proposed a WisCom framework to recognize the motivations behind why clients dislike specific mobile apps.

Gómez et al. [15] construct mobile apps reviews dataset to evolve a framework that identifies conceivably buggy mobile apps by enforcing a linkage in consent patterns and fault related reviews.

Chen et al. [16] presented a SimApp framework for identifying similar apps utilizing machine learning algorithms. SimApp inspects multimodal different data in app stores. They construct numerous kernel functions to degree app similarity. The outcomes exhibit that SimApp is powerful and promising for use in numerous applications, for example, app categorization, search and recommendation.

Tian et al. [5] research the main factors for recognizing high-rated apps by implementing random forest classifier. The test indicates that the main factors are promotional images numbers appeared on the app page, app size, and app version.

Lu et al. [17] suggest an approach to deal with classify mobile apps reviews automatically in light of non-functional requirements. They gathered 11,096 mobile apps reviews from Apple Store and Google Play.

Hameed et al. [11] explore existing Islamic apps accessible on Google Play app store. They handled the issue of the shortfall classification and the mis-categorization of Islamic apps. Therefore, they recommended another categorization for the Islamic apps' dependent on their common features such as download numbers, app ratings, and languages. They gathered proposed 5 distinct classes for the Islamic apps: Zakat, Qibla/Prayer Time, Quran, Hadith, and Supplications.

Abuelenin et al. [18] proposed hybrid system to find the most similar word in lexicon and increase the accuracy of Egyptian Arabic using the cosine similarity algorithm and the Information Science Research Institute Arabic stemmer (ISRI).

Al-Shamani et al. [12] construct Arb-AppsReview dataset for various research domains, such as gender detection, dialect analysis, sentiment analysis.

2) State-of-arts hybrid models: Heikal et al [19] propose a model which applies a hybrid model consists of CNN, and LSTM on ASTD. This model prediction performance is to 65%.

Al-Twairesh et al [20] suggest a model which applies a hybrid model SF+ GE + ASEH on SemEval. This model prediction performance is to 80.36%.

Mohammed et al. [21] propose a model which applies a hybrid model LSTM+Augmented on Arabic tweets. This model prediction performance is to 88.05%.

Furthermore, few previous works suggested a hybrid classification SA model for classify Egyptian Dialect Arabic mobile apps reviews.

III. A Hybrid Sentiment Analysis Classification Approach for Mobile Apps Arabic Slang Reiews (MASR) Methodology

This paper methodology depends on previous qualitative, quantitative and SLR research methodology [22]. It built according to previous observations after analyzing ASA survey, comparative framework [23] and future relationship hypothesis, user satisfaction surveys and case studies. The proposed methodology will be based on applying Natural Processing Language (NLP) and Data Mining (DM) Tools, Methods and Techniques. It depends on the quality of extracted features that express user opinion and its sentiment for Arabic Mobile Apps'. Finally, the main goal for it is to help developers improve and enhance new releases of Mobile Apps to meet rapidly changing in requirements evolution.

This research adopted a hybrid classification model which consist of six phases for collect, analyze and classify sentimental Arabic Dialect mobile apps reviews on google play store, as shown in Fig. 1.

This paper construct six phases for a hybrid classification Model methodology as indicated by Fig. 1; phase 1 MASR collection phase involves how to scrape and gather the dataset from google play store via Appbo[1] scraper tool and describing the dataset characteristics. The second phase involves the implementing of various pre-processing steps which will be applied on MASR dataset. The third phase is implementing feature extraction using Bag of Words (BOW) and Tf-idf. The fourth phase is implementing famous supervised machine learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), Linear Regression (LR), Neural Network (NN), and KNN classifier. The fifth phase proposing hybrid classification techniques according to the results of classifiers which accomplish highly accuracy results from the previous phase to enhance MASR accuracy results.

[1] https://appbot.co/

The last phase is to evaluate and compare the classification results utilizing recall, precision and accuracy.

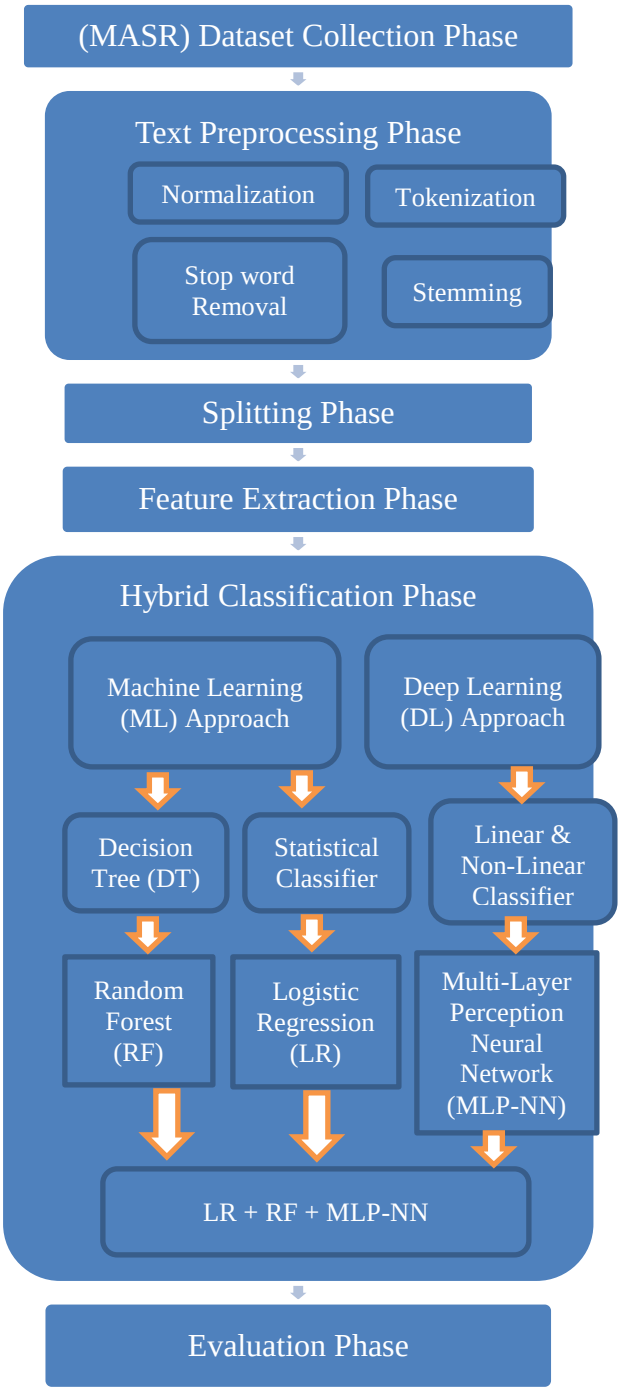

Fig. 1. HSACA-MASR Methodology Phases.

1) First Phase: Mobile Apps Arabic Slang Reviews (MASR) Dataset Collection Phase.

In this research, the Mobile Apps Egyptian DA reviews dataset was extracted using Appbot scraper tool which follows those steps:

- Choose Google Play Store.
- Select 9 various categories of mobile apps as shown in Table I.
- Focus on reviews of Egyptian mobile apps, and another Egyptian reviews for non-Egyptian mobile apps such as: Instagram, as shown in Table I.
- Save extracted attributes and reviews in CSV file: app category, app name, review, rating, and review polarity, as shown in Table II.

TABLE I. APP CATEGORY, APP NAME, APP RATING

	App category	*App name*	*App rating*
1	**Social**	Instagram[2]	4.4
2	**Lifetyle**	ContactCars[3]	4.5
3	**Travel & Locals**	Egypt Air[4]	4
4	**Shopping**	Kazyon[5]	4
		Olx Egypt[6]	4.3
5	**Tools**	Otlob[7]	4
		Shareit[8]	4.1
6	**Medical**	Vezeeta[9]	4.7
7	**Productivity**	Ana Vodafone[10]	4.2
8	**Education**	Aladwaa Education[11]	4
		بنك المعرفة المصرى [12]	4
9	**Maps & Navigation**	Careem[13]	4.2

2) MASR Properties: MASR dataset comprises of 2469 reviews made up of 653 positive, 756 neutral and 1060 negative reviews. A negative review is characterized as a review that has been given a rating of "1" or "2" or "3". A positive review is one where the review has been given a rating of "3" or "4" or "5". At last, Neutral reviews with a rating of "1" or "2" or "3" or "4" or "5". The MASR dataset was made from the gathered data and comprises of the following fundamental attributes as shown in Table II.

3) MASR distribution: MASR dataset covers 2469 mobile apps reviews contributed by various reviewers from 12 mobile apps which covers nine various mobile apps categories such as social, lifestyle, education, maps and navigation, productivity, shopping, travel and tools. The negative reviews comprise 43% of the absolute number of reviews when contrasted with

[2] https://play.google.com/store/apps/details?id=com.instagram.android
[3] https://play.google.com/store/apps/details?id=net.sarmady.contactcarswithtabs
[4] https://play.google.com/store/apps/details?id=com.linkdev.egyptair.app
[5] https://play.google.com/store/apps/details?id=com.inova.kazyon
[6] https://play.google.com/store/apps/details?id=com.olxmena.horizontal
[7] https://play.google.com/store/apps/details?id=com.semicoloneg.otlob
[8] https://play.google.com/store/apps/details?id=com.lenovo.anyshare.gps
[9] https://play.google.com/store/apps/details?id=com.ionicframework.vezeetapatientsmobile694843
[10] https://play.google.com/store/apps/details?id=com.emeint.android.myservices
[11] https://play.google.com/store/apps/details?id=com.nahdetmisr.adwaa
[12] https://play.google.com/store/apps/details?id=banke.elma3regypt
[13] https://play.google.com/store/apps/details?id=com.careem.acma

the 26% of the positive ones. Furthermore, 31% of the reviews are "neutral". As expected, the negative reviews are the greater part class. Fig. 2 presents the classification of ratings for our extracted dataset.

TABLE II. MASR DATASET ATTRIBUTES

Attribute	*Description*
Mobile App Category	Category of mobile app which include various category according to extracted datasets (Social, Lifestyle, Travel & Local, Shopping, Tools, Medical, Productivity, Education, or Maps & Navigation).
Mobile App Name	Name of selected Mobile App.
Review	opinion of reviewer's written in the ED which is mixing between MSA or DA.
Rating	Applies scale from 1 to 5 showing the scope of the reviewer's satisfaction. Positive reviews instead of using the previous scale from 1 to 10.
Review Polarity	Denotes the sentiment of the review with "+1" for a positive review, "−1" for a negative review, and "0" for a neutral review.

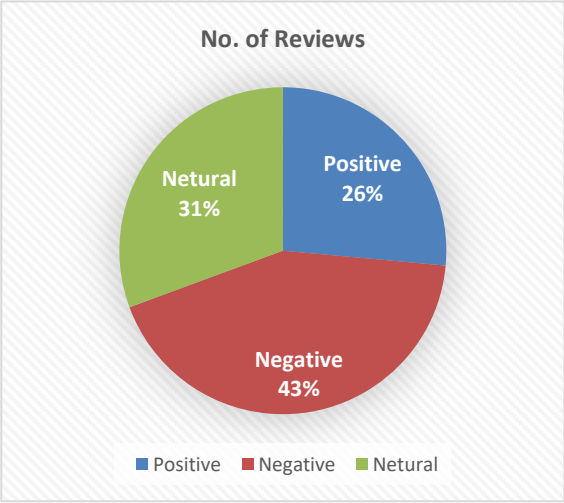

Fig. 2. MASR Dataset Polarity Distribution.

Table III represents samples of MASR datasets which contains app category, app name, review, translated review, rating, and polarity (Negative, Positive, Neutral).

4) Second phase: Text Preprocessing phase: The initial step is to implement text pre-processing so as to evolve the performance of classifiers by changing the text into a format as suitable as possible. To achieve this, many stages are executed; specifically, normalization, tokenization, stop-word removal and stemming.

5) Normalization: This stage includes the accompanying steps: Remove punctuation marks and special characters, remove tatweel kashida symbol ("--"), remove of all diacritics, remove digit numbers (0-9), remove repeated characters, remove all non-Arabic words, replace each final letter (ي) with (ى), replace initial letter alef-hamza (أ، إ،ء،ئ،ؤ، آ) with (ا), and replace each final letter (ة) with (ه).

TABLE III. MASR REVIEWS

App Category	*App Name*	*Review*	*Translated Review*	*Rating*	*Polarity*
Productivity	Ana Vodafone	معرفتش ادفع ببطاقة الائتمان	I don't pay by credit card	3	Negative
Travel & Local	EGYPTAIR	برنامج رائع برجاء اضافة صاله السفر والوصول في حالة الرحلة	Wonderful program, please add the travel and arrival hall in case of flight	4	Positive
Tools	SHAREit	طبيق ممتاااز بس احيانا لما احاول ابعت حاجه لإصدار اقل مني ميبعتش	An excellent application, but sometimes when I try to send something less than me, I need to issue it	5	Neutral

6) Tokenization: For author/s of more than two affiliations: To change the default, adjust the template as follows. By tokenizing, you can appropriately separate text by word or by sentence. This will permit act with smaller sets of text that are still comparatively meaningful regular outgoing of the context of the remainder of the text. In this research, Regexp Nltk14 method applies on MASR dataset. It divides a string into substrings utilizing a standard expression. It can utilize its regexp to look like delimiters instead.

7) Stop word removal: The second stage is to eliminate all stop-words from the reviews. Stop words are characterized as words that don't increase any sentiment value to a review; they are typically the most widely recognized words in a language. They can either be specially made or gained from the web. Unfortunately, there is no clear list accessible and there are slight lists accessible for the Arabic language. This research adjusted Arabic stopword list from many resources in addition to Egyptian stopword list from [24].

8) Stemming: Stemming is a text processing method of decreasing a word to its root. It maps various patterns of the similar word to a public "stem" - for example, the Arabic stemmer maps طفل, اطفال, الاطفال, اطفالكم, فأطفالكم, اطفالهم, والاطفال, والطفولة, وطفل, فاطفالهم, and طفلتان to طفل. In this research, Snowball[15] stemmer applies on MASR dataset.

a) Third phase: Splitting phase: MASR dataset was separated into two sections: training sets, and testing sets. The training sets represent 70% of the datasets, and the testing sets represents 30%. The training sets utilized to train models, while the testing sets utilized to evaluate models.

[14] https://www.nltk.org/_modules/nltk/tokenize/regexp.html
[15] https://git.texta.ee/texta/snowball/-/blob/master/python/testapp.py

b) Fourth phase: Feature extraction phase to estimate classifiers performance, this research utilized various variety of features. Those features can be Bag-of-Words (BOW) with TF-IDF (Term Frequency Inverse Document Frequency).

9) Bag of Words (BOW) [16]: BOW is a process of eliciting features from text for utilize in modeling, such as with ML algorithms. BOW model assigns a corpus with word counts for every document.

*10)*Term Frequency- Inverse Document Frequency (TF-IDF) [17]: Tf-IDF weight is a statistical measure utilized to estimate how significant a word is to a document in a corpus. The significance grows proportionally to the frequency of times a word represents in the document. It is formed by two sections:

$$Tfidf = \log(word, review) * \log \frac{\sum reviews}{\sum freq\ of\ words} \quad (1)$$

a) Fifth phase: Hybrid supervised classification approach phase: This phase performs two subsections: the first issue is applying ML approach which performs five selected ML classifiers which utilized extensively for ASA: Logistic Regression (LR) [25] [26] [27] [28], Naïve Bayes (NB) [29] [30] [31] [32], K-Nearest Neighbors (KNN) [33] [34] [31] [35], Random Forest (RF) [36] [37] [38] and SVM [33] [39] [40] in addition applying DL approach which performs DL classifier Multi-Layer Perceptron Neural Network (MLP-NN) which applied in [36] [37] for ASA.

For the second issue: This research intends to propose a novel Hybrid Supervised Classification Approach to automatically classify and predict the polarity of mobile apps Arabic Slang user reviews. This model mixes various supervised ML, and DL approaches. In ML approach, we suggest various modeling approaches: decision tree approach, and statistical approach. While in DL approach, we suggest linear & non-linear approach. In decision tree approach, we apply RF classifier. In Linear & Non-Linear approach, we apply MLP-NN classifier. In Statistical approach, we apply LR classifier. The reason for selecting those classifiers came after applying various ML classifiers in a previous phase. The results shows that the top classifiers that gain best accuracy for classify or predict MASR datasets are: RF, LR, and MLP-NN. Finally, we propose to apply a hybrid classification model that combines those three techniques to improve accuracy performance.

b) Six phase evaluation phase: To evaluate ML, DL, and our proposed hybrid classification approaches algorithms, this research applied 10-fold cross validation. This paper assessed performance of those models utilizing various evaluation measures: Accuracy (ACC) [41], F-measure [41], Precision (PRE) [41], Recall (REC) [41], Area Under the Curve (AUC) [42], and Ensemble classifier average [28].

$$Accuracy = \frac{\sum TP+\sum TN}{\sum TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F-Measure = \frac{2*Precision*Recall}{(Precision*Recall)} \quad (5)$$

$$F\ abs = \left(\frac{AUC\ non-iv}{AUC\ iv}\right) * \left(\frac{Dose\ iv}{Dose\ non-iv}\right) \quad (6)$$

$$Ensemble(AVG) = \sum \frac{1}{n}(a1 + a2 + \cdots . + an) \quad (7)$$

IV. Results and Discussion

For empirical study, ORANGE Data Mining tool utilizes a component-based, inclusive model for DM and ML users and developers. Also, this research utilizes it for ML, and DL Models purposes. It is a combination of Python-based, and NLTK library modules which perform a set of functions such as data input, pre-processing, splitting, visualization, classification, prediction, and evaluation. Classifier methods used to classify MASR dataset utilizing: ML approach which perform KNN, SVM, NB, & LR, DL approach which perform MLP-NN for ASA. In addition, this paper suggests a novel hybrid classification technique which combined from two top ML classifiers in addition to DL classifier: LR + RF +MLP-NN to enhance accuracy for classification and prediction. k-fold cross-validation was utilized with k = 10. Accuracy, F1, Precision, Recall, AUC were utilized for evaluate MASR sentiment polarity datasets.

The results are discussed separately for each evaluation criterion. Moreover, to ensure the performance of the classifiers, this paper combined various domains to test the accuracy of various ML, DL, and our proposed hybrid approach using Arabic dialect features.

1) Accuracy (AUC): Fig. 3 represents the performance of three various classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

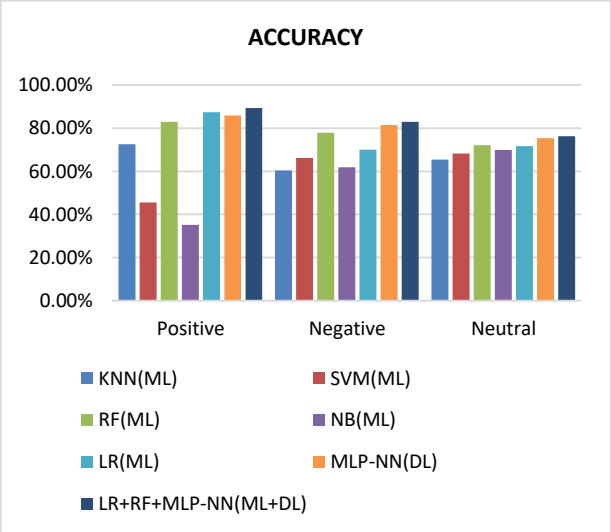

Fig. 3. Accuracy of Ml, Dl, and Hybrid (ML+DL) Approach.

[16]https://gist.github.com/mwitiderrick/363a71bc0d686383a33132aa9f896fce

[17] https://scikit-learn.org/stable/modules/feature_extraction.html

After applying ML classifiers on Positive Sentiments, results show that LR (87.5%), and RF (83%) shows better accuracy compared to a KNN (72.6%), SVM (45.5%), and NB (35.1%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (86%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (89.4%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that RF (78%), and LR (70%) shows better accuracy compared to a SVM (66.2%), KNN (60.4%), and NB (61.9%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (81.5%) perform better accuracy than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (83%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that RF (72.1%), and LR (71.7%) shows better accuracy compared to a NB (69.9%), SVM (68.3%), and KNN (65.5%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (75.4%) perform better accuracy than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (76.3%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

2) Precision (PRE): Fig. 4 represents the various precision results of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

After applying ML classifiers on Positive Sentiments, results mention that LR (91%) and RF (66.3%) shows better Precision results compared to a KNN (48.8%), SVM (31.8%), and NB (29%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Precision of LR (91%) perform better than Precision of NLP-NN (70.8%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Precision result is (84.4%). So, LR performs better Precision result than our proposed hybrid approach.

After applying ML classifiers on Negative Sentiments, results mention that NB (96.3%) and SVM (73.9%) shows better Precision results compared to a RF (71.8%), LR (59.6%), and KNN (54.1%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Precision of NB (96.3%) perform better than Precision of MLP-NN (76.1%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Precision result is (79.1%). So, NB performs better Precision results than our proposed hybrid approach, and MLP-NN.

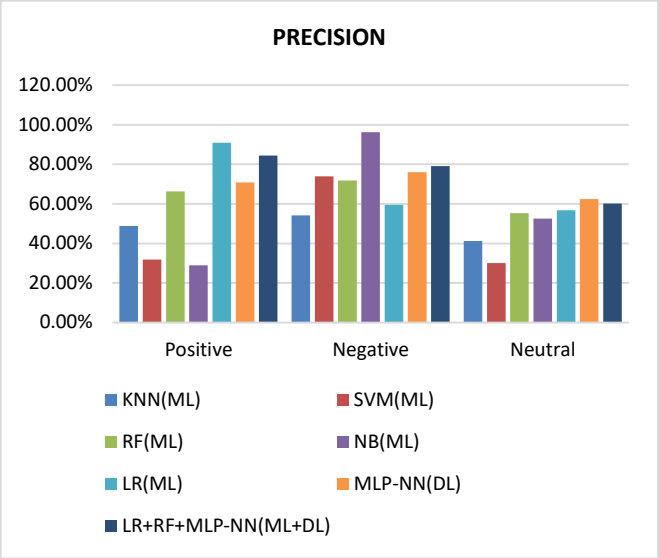

Fig. 4. Precision of Ml, Dl, and Hybrid (ML+DL) Approach.

After applying ML classifiers on Neutral Sentiments, results mention that LR (56.8%) and RF (55.3%) shows better Precision results compared to a NB (52.6%), KNN (41.3%), and SVM (30.1%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Precision of NLP-NN (62.5%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Precision result is (60.2%). So, MLP-NN(DL) performs better Precision results than our proposed hybrid approach.

3) Recall (REC): Fig. 5 illustrates the various recall results of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

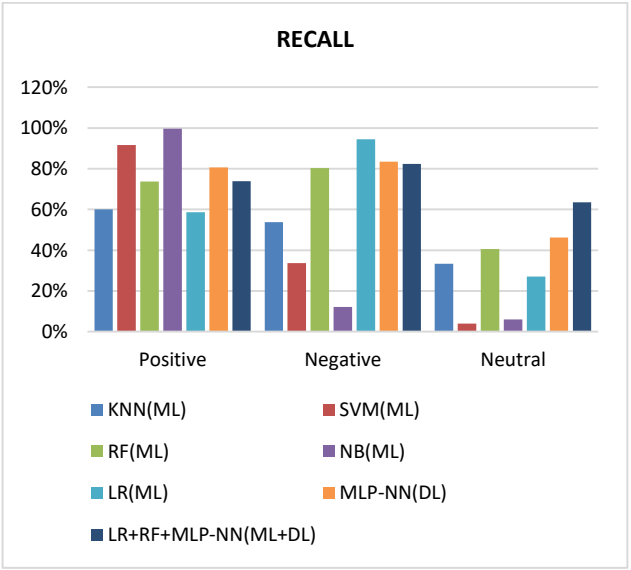

Fig. 5. Recall of Ml, Dl, and Hybrid (ML+DL) Approach.

After applying ML classifiers on Positive Sentiments, results mention that NB (99.7%), and SVM (91.6%) shows better Recall results compared to a RF (73.7%), KNN (60%), and LR (58.7%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Recall of MLP-NN (80.6%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Recall of is (73.8%). So, NB and SVM perform better recall results than DL (MLP-NN) and our proposed hybrid approach (LR+RF+MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that LR (94.5%) and RF (80.3%) shows better Recall results compared to a KNN (53.8%), SVM (33.6%), and NB (12.2%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Recall of MLP-NN (83.5%). And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that Recall of is (82.3%). So, LR performs better recall results than DL (MLP-NN), our proposed hybrid approach (LR+RF+MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that) RF (40%) and KNN (33.3%) shows better Recall results compared to a LR (27.1%), NB (6%), and SVM (4%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that Recall of NLP-NN (46.2%) perform better recall than top two ML classifiers RF, KNN. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (63.5%) than the top three classifiers: ML (RF, KNN), and DL (MLP-NN).

4) F1-Measure: Fig. 6 represents the performance of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

After applying ML classifiers on Positive Sentiments, results mention that LR (71.3%) and RF (69.8%) shows better F1-Measure results compared to a KNN (53.8%), SVM (47.2%), and NB (45%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (75.4%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (78.8%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that RF (75%), and LR (73.1%) shows better F1-Measure results compared to a KNN (54%), SVM (46.2%), and NB (21.6%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (79.6%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (80.1%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that RF (46.8%), KNN (36.9%) and LR (36.7%) shows better F1-Measure results compared to a SVM (6.7%), and NB (10%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (53.1%) perform better than top two ML classifiers LR, RF. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (61.8%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

5) Area Under the Curve (AUC): Fig. 7 represents a graph of the various AUC results of three different classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN).

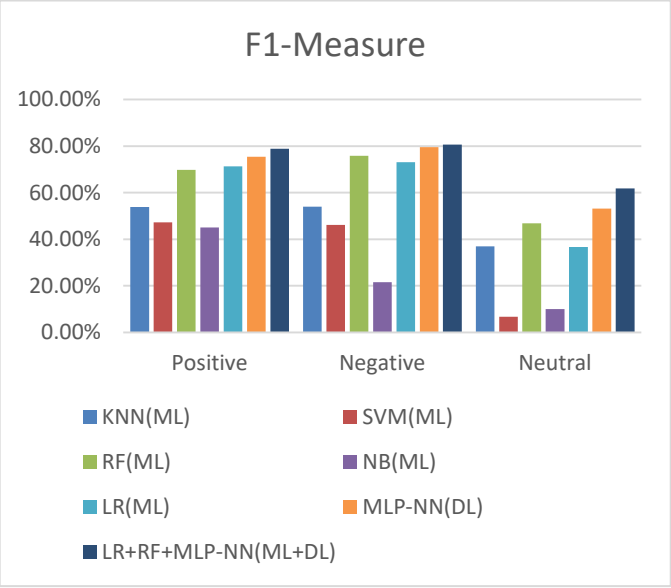

Fig. 6. F1-Measure of Ml, Dl, and Hybrid (ML+DL) Approach.

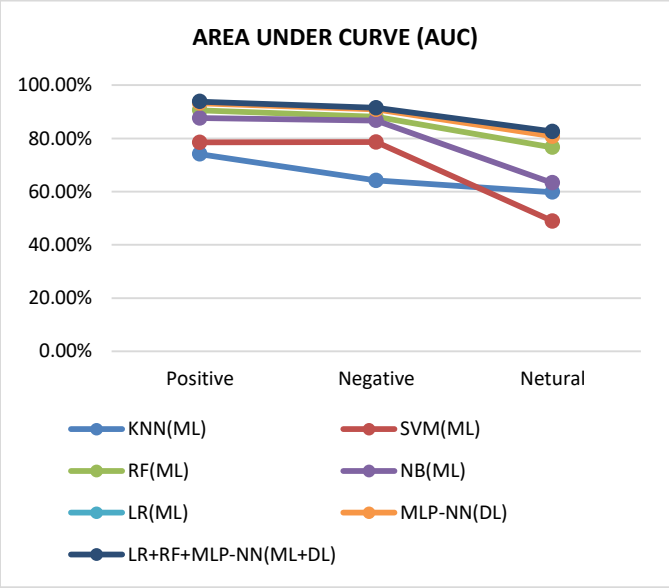

Fig. 7. AUC of Ml, Dl, and Hybrid (ML+DL) Approach.

After applying ML classifiers on Positive Sentiments, results mention that LR (93.5%) and RF (90.6%) shows better AUC results compared to a NB (87.6%), SVM (78.5%), and KNN (74.1%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of MLP-NN (93.22%) is approximate to ML classifier LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (93.8%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Negative Sentiments, results mention that LR (91.1%) and RF (88.1%) shows better AUC results compared to a NB (86.7%), SVM (78.6%), and KNN (64.2%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of NLP-NN (90.9%) is approximate to ML classifier LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (91.5%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

After applying ML classifiers on Neutral Sentiments, results mention that LR (81.7%) and RF (76.6%) shows better AUC results compared to a NB (63.3%), KNN (59.8%), and SVM (48.9%), and respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of NLP-NN (80.8%) is approximate to ML classifier LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (82.6%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

6) Ensemble classifier averaging: Fig. 8 represents the average performance of three various classification approaches: ML classifiers (KNN, SVM, NB, RF, LR), DL classifier (MLP-NN), and our proposed hybrid classification model approach: ML+DL (LR+RF+MLP-NN) utilizing various evaluation criteria.

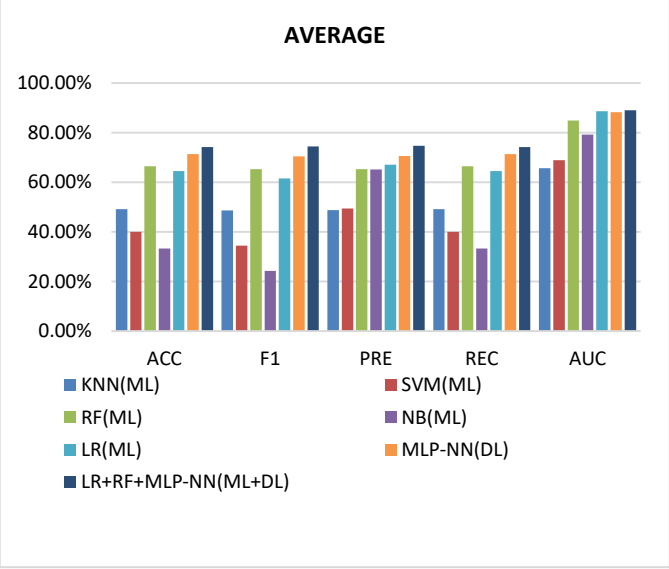

Fig. 8. Average of Ml, Dl, Hybrid (ML+DL) Approach and Evaluation Metrics.

Accuracy. After applying ML classifiers, results mention that LR (72%), and RF (70%) shows better accuracy compared to a SVM (50%), KNN (49%), and NB (45%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that NLP-NN accuracy (69%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better accuracy (74.2%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

Precision. After applying ML classifiers, results mention that LR (71.8%), and RF (69.6%) shows better precision results compared to a NB (62.6%), SVM (53.9%), and KNN (51.4%) respectively. In addition, after applying DL classifier: MLP-NN, results observe that precision of NLP-NN (68.2%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better precision results (74.9%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

Recall. After applying ML classifiers, results mention that LR (72.3%), and RF (70.3%) shows better recall results compared to a SVM (50.1%), KNN (49%), and NB (45.2%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that the recall of NLP-NN (69.1%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better recall results (74.2%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

F1-Measure. After applying ML classifiers, results mention that LR (71.8%), and RF (69.6%) shows better F1-Measure results compared to a KNN (47.9%), SVM (45.7%), and NB (41.8%), respectively. In addition, after applying DL classifier: MLP-NN, results observe that F1-Measure of NLP-NN (68.2%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better F1-Measure results (74.2%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

AUC. After applying ML classifiers, results mention that LR (88.9%), and RF (872%) shows better result of AUC compared to a NB (829%), SVM (72.8%), and KNN (70%) respectively. In addition, after applying DL classifier: MLP-NN, results observe that AUC of NLP-NN (86.1%) is approximate to ML classifiers: RF, and LR. And finally, after applying our proposed approach ML+DL (LR+RF+MLP-NN), results recognize that it performs better AUC (89.6%) than the top three classifiers: ML (LR, RF), and DL (MLP-NN).

Finally, researchers summarize our performance for a proposed hybrid (LR+RF+MLP-NN) approach results as follow:

- Positive polarity: performs higher performance results in the following evaluation criteria: ACC (89.4%), F1 (78.8%), and AUC (93.8%).

- Negative polarity: performs higher performance results in the following evaluation criteria: ACC (83%), F1 (80.7%), and AUC (91.5%).
- Neutral polarity: performs higher performance results in the following evaluation criteria: ACC (76.3%), REC (63.5%), F1 (61.8%), and AUC (82.6%).
- Average: performs higher performance results in the following evaluation criteria: ACC (74.3%), PRE (74.8%), REC (74.3%), F1 (74.5%), AUC (89.1%).

TABLE IV. COMPARISON BETWEEN STATE-OF-ARTS HYBRID MODELS AND OUR HYBRID MODEL

Study	Dataset	Hybrid Models	Accuracy
Heikal et al. [19]	ASTD	CNN + LSTM	65.05%
Al-Twairesh et al. [20]	SemEval	SF+ GE + ASEH	80.36%
Al-Azani et al. [43]	ASTD	SGD + SGD + NuSVC	85.28%
Basir et al. [44]	COVID	CNN+ BiGRU + FastText	85.4%
Saleh et al. [38]	AJGT	LR+CBOW	86.11%
Mohammed et al. [21]	Arabic tweets	LSTM+Augmented	88.05%
Our Hybrid Approach	MASR	LR+RF+MLP-NN	89.4%

In Table IV, a comparison between the performance of our model accuracy and state-of-arts hybrid models on the various Arabic datasets (SemEval, ASTD, COVID datasets, AJGT) is presented. Researchers observe the excellence of our proposed hybrid model approach compared to the previous works.

V. CONCLUSION

This paper aims to collect a simple dataset of Mobile Apps Arabic Slang Reviews (MASR) which focus on Egyptian Arabic Slang for sentiment analysis purposes. In addition, propose a hybrid supervised classification approach which combine ML, and DL approaches to automatically predict user requirements evolution to help developers update new versions. In ML approach, apply a LR which considered a statistical method, and RF which considered a decision tree method. In DL approach, apply MLP-NN which considered a linear and non-linear method. This paper utilized various evaluation metrics like: accuracy, f-measure, recall, precision, AUC, and ensemble classifier averaging. Results show that our proposed hybrid supervised classification approach achieves good performance results in the following:

- In Positive polarity, ACC (89.4%), F1 (78.8%), and AUC (93.8%).
- In Negative polarity, ACC (83%), F1 (80.7%), and AUC (91.5%).
- In Neutral polarity, ACC (76.3%), REC (63.5%), F1 (61.8%), and AUC (82.6%).
- In Average, ACC (74.3%), PRE (74.8%), REC (74.3%), F1 (74.5%), AUC (89.1%).

A limitation in this research is the size of the dataset because it focuses only on Egyptian Arabic Slang mobile reviews. However, it considered a contribution because till now no studies concentrate on it.

VI. FUTURE WORK

In future, researchers intend to accomplish various researches in various points:

1) Apply our proposed hybrid supervised approach for automatically classify Mobile Apps categories.

2) Apply our proposed hybrid supervised approach for different Mobile Apps Arabic Slang datasets in different languages.

3) Add different feature extraction methods like word embedding, and word enrichment and n-grams, also apply different tokenization, and stemming methods.

4) Propose different hybrid ML, and DL modelling approaches and compare them with our proposed approach on different Arabic Slang datasets.

5) Apply also lexicon approach in addition to MASR dataset.

6) Extract functional, and Non-Functional, and Sentimental requirements from MASR datasets using Topic Modeling approach.

REFERENCES

[1] Fuad and M. Al-Yahya, "Analysis and Classification of Mobile Apps Using Topic Modeling: A Case Study on Google Play Arabic Apps.," Complexity, vol. 2021, 2021.

[2] I. Malavolta, S. Ruberto, T. Soru and V. Terragni, "Hybrid mobile apps in the google play store: an exploratory investigation," in Proceedings of the 2nd ACM International Conference on Mobile Software Engineering and Systems, Florence, Italy, 2015.

[3] G. Berardi, A. Esuli, T. Fagni and F. Sebastiani, "Multi-store metadata-based supervised mobile app classification," in In Proceedings of the 30th Annual ACM Symposium on Applied Computing, 2015.

[4] K. Elshakankery and M. F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis," Egyptian Informatics Journal, vol. 20, no. 3, pp. 163-171, 2019.

[5] Y. Tian, M. Nagappan, D. Lo and A. E. Hassan., "What are the characteristics of high-rated apps? a case study on free android applications," in In Proceedings of the 2015 IEEE international conference on software maintenance and evolution (ICSME), 2015.

[6] W. Martin, App store analysis for software engineering, UK, , London: University College London, 2017.

[7] A. Finkelstein, M. Harman, Y. Jia, W. Martin, F. Sarro and Y. Zhang, "Investigating the relationship between price, rating, and popularity in the Blackberry World App Store," Information and Software Technology, vol. 87, pp. 119-139.

[8] A. Finkelstein, M. Harman, Y. Jia, . F. Sarro and Y. Zhang, "Mining App Stores: Extracting Technical, Business and Customer Rating Information for Analysis and Prediction," Research Note RN/13/21, 2013.

[9] E.-Y. Jung, C. Baek and &. J.-D. Lee, "Product survival analysis for the App Store," Marketing Letters, vol. 23, no. 4, pp. 929-941, 2012.

[10] M. Harman, Y. Jia and Y. Zhang, "App store mining and analysis: MSR for app stores," in In Proceedings of the 2012 9th IEEE working conference on mining software repositories (MSR), 2012.

[11] A. Hameed, H. A. Ahmed and N. Z. Bawany, "Survey, analysis and issues of Islamic Android apps," Elkawnie: Journal of Islamic Science and Technology, vol. 5, no. 1, pp. 1-15, 2019.

[12] M. Al-Shamani, M. Al-Sarem, F. Saeed and W. Almutairi, "Designing an Arabic Google Play Store User Review Dataset for Detecting App

Requirement Issues," in In Advances on Smart and Soft Computing, Singapore, Springer, 2022, pp. 133-143.

[13] S. R. El-Beltagy and A. Ali, "Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study," in Proceedings of the 9th International Conference on Innovations in Information Technology (IIT), 2013.

[14] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong and N. Sadeh, "Why people hate your app: making sense of user feedback in a mobile app store," in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 2013.

[15] M. Gómez, R. Rouvoy, M. Monperrus and L. Seinturier, "A Recommender System of Buggy App Checkers for App Store Moderators," in Gomez, Maria, et al. "A recommender system of buggy app checkers for app store moderators." 2015 2nd ACM International Conference on Mobile Software Engineering and Systems., 2015.

[16] N. Chen, S. CH Hoi, S. Li and X. Xiao, "SimApp: A framework for detecting similar mobile applications by online kernel learning," in In Proceedings of the eighth ACM international conference on web search and data mining, 2015.

[17] M. LU, and P. LIANG, "Automatic classification of non-functional requirements from augmented app user reviews," in Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, 2017.

[18] S. Abuelenin, S. Elmougy and E. Naguib, "Twitter sentiment analysis for arabic tweets," in Proceedings of International conference on advanced intelligent systems and informatics, Cham, , 2017.

[19] M. Heikal, M. Torki and . N. El-Makky, "Sentiment analysis of Arabic tweets using deep learning," Procedia Computer Science, vol. 142, pp. 114-122, 2018.

[20] N. Al-Twairesh and H. AL-Negheimish, "Surface and Deep Features Ensemble for Sentiment Analysis of Arabic Tweets," IEEE Access, vol. 7, pp. 84122-84131, 2019.

[21] A. Mohammed and R. Kora, "Deep learning approaches for Arabic sentiment analysis," Social Network Analysis and Mining, vol. 9, no. 1, pp. 1-12, 2019.

[22] R. E. Saudy, E. S. Nasr, A. E. D. M. El-Ghazaly and M. H. Gheith, "Use of Arabic sentiment analysis for mobile applications' requirements evolution: trends and challenges," in Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cham, 2017.

[23] R. E. Saudy, E. S. Nasr, A. E. D. M. El-Ghazly and M. H. Gheith, "A Comparative Framework for Arabic Sentiment Analysis Research," in The 54thAnnual Conference on Statistics, Computer Sciences and Operation Research, Egypt, 2019.

[24] W. Medhat, A. Yousef and H. Korashy, "Egyptian dialect stopword list generation from social network data," The Egyptian Journal of Language Engineering, vol. 2, no. 1, pp. 43-55, 2015.

[25] M. M. Al-Tahrawi, "Arabic Text Categorization Using Logistic Regression," International Journal of Intelligent Systems and Applications, vol. 7, no. 6, p. 71, 2015.

[26] M. Al-Omari, "logistic regression optimisation for Arabic customers' reviews," International Journal of Business Intelligence and Data Mining, vol. 20, no. 3, pp. 251-273, 2022.

[27] R. Ismail, M. Omer, M. Tabir, N. Mahadi and I. Amin, "Sentiment analysis for Arabic dialect using supervised learning," in In Proceedings of the International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018.

[28] A. Hawalah, "A Framework for Arabic Sentiment Analysis Using Machine Learning Classifiers," Journal of Theoretical and Applied Information Technology, vol. 97, no. 17, pp. 4478-4489, 2019.

[29] J. O. Atoum and M. Nouman, "Sentiment analysis of Arabic Jordanian dialect tweets," International Journal of Advanced Computer Science and Applications, vol. 10, no. 2, pp. 256-262, 2019.

[30] A. Alnawas and A. Nursal , "Sentiment analysis of Iraqi Arabic dialect on Facebook based on distributed representations of documents," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 18, no. 3, pp. 1-17, 2019.

[31] R. M. Duwairi and I. Qarqaz, "A framework for Arabic sentiment analysis using supervised classification.," International Journal of Data Mining, Modelling and Management, vol. 8, no. 4, pp. 369-381, 2016.

[32] M. Alassaf and A. M. Qamar, "Improving sentiment analysis of Arabic tweets by One-way ANOVA," Journal of King Saud University-Computer and Information Sciences, vol. 1, no. 0, pp. 1-11, 2020.

[33] A. S. AL-Jumaili, "A hybrid method of linguistic and statistical features for Arabic sentiment analysis," Baghdad Science Journal, vol. 17, no. 1, 2020.

[34] A. K. Al-Tamimi, A. Shatnawi and . E. Bani-Issa, "Arabic sentiment analysis of YouTube comments," in In Proceedings of the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Aqaba- Jordan, 2017.

[35] M. E. M. Abo, N. Idris, R. Mahmud, A. Qazi, A. T. H. Ibrahim , J. Zubairu Maitama, U. Naseem, S. K. Khan and S. Yang, "A Multi-Criteria Approach for Arabic Dialect Sentiment Analysis for Online Reviews: Exploiting Optimal Machine Learning Algorithm Selection," Sustainability, vol. 13, no. 18, pp. 1-20, 2021.

[36] S. Bessou and R. Aberkane, "Subjective Sentiment Analysis for Arabic Newswire Comments," Journal of Digital Information Management (JDIM), vol. 17, no. 5, pp. 289-295, 2019.

[37] A. A. Sayed., E. Elgeldawi, Z. M. Alaa and G. R. Ahmed, "Sentiment Analysis for Arabic Reviews using Machine Learning Classification Algorithms," in In Proceedings of the International Conference on Innovative Trends in Communication and Computer Engineering (ITCE, 2020.

[38] H. Saleh, S. Mostafa , . A. Alharbi, . S. El-Sappagh and T. Alkhalifah, "Heterogeneous Ensemble Deep Learning Model for Enhanced Arabic Sentiment Analysis," Sensors, vol. 22, pp. 1-28, 2022.

[39] S. Alhumoud, "Arabic sentiment analysis using deep learning for covid-19 twitter data," International Journal of Computer Science and Network Security, vol. 20, no. 9, pp. 132-138, 2020.

[40] A. Elhawil, Y. Trabelsi and M. Mahfoud, "Comparison between the NB and SVM methods for multiclass Arabic sentiment analysis," in In Proceedings of the IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA, 2021.

[41] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, , informedness, markedness and correlation," arXiv preprint arXiv:2010, vol. 16061, pp. 37-63, 2020.

[42] B. Yamout, Z. Issa, A. Herlopian, M. El Bejjani, A. Khalifa, A. S. Ghadieh and R. H. Habib, "Predictors of quality of life among multiple sclerosis patients: a comprehensive analysis," European Journal of Neurology, vol. 20, no. 5, pp. 756-764, 2013.

[43] S. Al-Azani and E.-S. M. El-Alfy, "Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text," Procedia Computer Science 109C, p. 359–366, 2017.

[44] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets," Knowledge-Based Systems, vol. 228, 2021.