Enhancing Business Intelligence with Hybrid Transformers and Automated Annotation for Arabic Sentiment Analysis

Wael M.S. Yafooz
Computer Science Department, College of Computer Science and Engineering, Taibah University, Medina, 42353, Saudi Arabia

Abstract—**Business is a key focus for many individuals, companies, countries and organisations. One effective way to enhance business performance is by analysing customer opinions through sentiment analysis. This technique offers valuable insights, known as business intelligence, which directly benefits business owners by informing their decisions and strategies. Substantial attention has been given to business intelligence through proposed machine learning approaches, deep learning models and approaches utilizing natural language processing methods. However, building a robust model to detect and identify users' opinion and automated text annotation, particularly for the Arabic language, still faces many challenges. Thus, this study aims to propose a hybrid transfer learning model that uses transformers to identify positive and negative user comments that are related to business. This model consists of three pretrained models, namely, AraBERT, ArabicBERT, and XLM-RoBERTa. In addition, this study proposes a hybrid automatic Arabic annotation method based on CAMelBERT, TextBlob and Farasa to automatically classify user comments. A novel dataset, which is collected from user-generated comments (i.e. reviews on mobile apps), is introduced. This dataset is annotated twice using the proposed method and human-based annotation. Then, several experiments are conducted to evaluate the performance of the proposed model and the proposed annotation method. Experiment results show that the proposed hybrid model outperforms the baseline models, and the proposed annotation method achieves high accuracy, which is close to human-based annotation.**

Keywords—Business intelligence; machine learning; sentiment analysis; transformers; BERT; Arabic annotation

I. Introduction

Business intelligence can be enhanced through Sentiment Analysis (SA), which provides a nuanced understanding of customer opinions, market trends and brand reputation. SA allows businesses to gauge public sentiment accurately by analysing large amounts of text data from sources, such as social media, customer reviews and feedback forms. This insight helps companies identify consumer impressions, pain points and emerging trends, allowing them to tailor their products, services and marketing strategies [1, 2]. In this manner, businesses can stay competitive by responding quickly to changes in consumer behaviour and market conditions by understanding sentiment.

By integrating SA with business intelligence systems, companies can make more informed and strategic decisions based on gained insights into the emotional tone of customer feedback, which provides context to quantitative data, such as sales figures and retention rates. Social media platforms, such as Facebook, Twitter and Instagram, have billions of active users who regularly share their thoughts, opinions and experiences. By analysing social media posts, comments and reviews, businesses can identify emerging trends, monitor brand sentiment, gather feedback on products and services and identify early potential issues and address them proactively [3,4]. Similarly, mobile app reviews provide businesses with valuable feedback from users regarding their experiences with the app. These reviews often contain rich insights into user preferences, pain points and suggestions for improvement. By analysing app reviews, businesses can identify recurring issues, prioritise feature enhancements and enhance user satisfaction. Today's digital age has created a marketplace where consumer opinion is crucial to maintaining a positive brand image. SA makes it easier for businesses to keep an eye on public perception and respond quickly to changes in public opinion.

Therefore, scholarly efforts have been made to use SA in many domains, such as healthcare [5–7], education [8–10] marketing [11–13], business [14–16] and finance [17, 18] with binary classification or multi classification. The main two challenges that this task suffers from are dataset and building mode [19, 20]. Firstly, in the dataset preparation, the labelling process, which is known as the annotation process, is important because it is related to the model's performance in terms of accuracy [21]. This process is time consuming and requires considerable effort from annotators to classify data to relevant classes. Secondly, a robust model to distinguish between the good and bad comments or words is required. Researchers have attempted to address this issue by using Machine Learning (ML) classifiers, Deep Learning (DL) models and Natural Language Processing (NLP) methods for many languages, such as English, French and Chinese. However, the Arabic language and its dialects lack scholarly attention due to the Arabic language having morphological richness with 22 countries with different dialects. Some researchers introduced methods for Arabic annotations, such as AraSenCorpus [19], Sentialg [22], Arasenti-tweet [23] and ZAEBUC [24]. The majority of SA works in the literature review is based on manual annotation by Arabic native speakers [25–29] or by automatic annotation tools [22, 30–34]. Some of them use Google Translator to translate from Arabic to English, and then they apply automatic annotation tools [24]. In such manner, Google deals with modern standard Arabic, not Arabic dialects. Therefore, this still remains a challenge in the Arabic language due to its complex morphology, dialectal variation, ambiguity and contextual understanding.

Therefore, the purpose of this study is to propose a robust model to detect the users' sentiment to improve business intelligence and to propose methods for automatic Arabic annotation. The proposed model is a hybrid transfer learning model designed to distinguish between the good and bad user comments; it consists of XLM-RoBERTa, AraBERT Ver2 and Arabic BERT models. These models are all based on Bidirectional Encoder Representations from Transformers (BERT) which are Transformers architectures. In addition, this hybrid model comprises two methods which have been used; the first being the voting mechanism, and the second being the fusion feature. In the automatic Arabic annotation method, three annotator tools, i.e. CAMeLBERT, TextBlob and Farasa, are utilised. This method is called the Artificial Intelligence Annotator (AIA). Additionally, a novel dataset that has been collected from mobile apps, which consists of 223,341 user-generated comments, is introduced to validate the performance of the proposed model and to determine how well the AIA method accurately assigns the user comments to the relevant classes. The same dataset is also annotated by three Arabic native speakers; this process is called Human-based Annotation (HA). Then, several NLP methods, such as data cleaning, preprocessing and data annotation, are applied on the dataset to validate the proposed hybrid transfer learning model and the AIA. Several experiments are conducted using ML classifiers, DL models and transformers. The results of these experiments were in terms of the most common metrics, which are, recall, precision, F1-score, Area Under Curve (AUC)–Receiver Operating Characteristic Curve (ROC) and accuracy. Experiment results show that the proposed model is based on the fusion features and voting mechanism which outperformed all the baselines in terms of accuracy, i.e. 97.24% and 98.11%, respectively. Additionally, the experiment results show the AIA method is closely tied to HA in Arabic corpora annotation.

The contributions of this study are multifaceted and can be summarised as follows:

- Proposed a hybrid model of transformers (transfer learning) to detect business SA based on mobile apps that are related to home delivery and taxi. This model consists of XLM-RoBERTa, AraBERT Ver2 and ArabicBERT. In this model, two methods are used: the voting mechanism and the fusion feature.

- Proposed a hybrid automatic Arabic annotation method for the Arabic corpora, which consists of the CAMelBERT, TextBlob and Farasa methods.

- Introduced a novel Arabic dataset which consists of 223,341 user-generated comments collected from reviews of mobile apps. It was reduced after annotation process to 63,313 user comments using AIA method and 54,988 using the HA method.

- Compared the model performance in terms of accuracy with various ML classifiers, DL models, transformers for AIA and HA methods.

The remainder of this paper is organized as follows. Section II provides a review of recent studies that focus on sentiment analysis and business intelligence. Section III explains the methods used to achieve the objectives of this study. The results and experiments are presented in Section IV. Discussion is given in Section V. Finally, the paper concludes in Section VI, summarizing the findings and implications of the study.

II. Related Studies

This section explores the studies that focus on sentiment analysis from a business intelligence perspective and also focuses on the Arabic annotation methods for the Arabic corpora.

In business intelligence and social media, Kurnia & Suharjito [1], developed a business intelligence dashboard to observe the performance of each topic or channel of news posted on social media apps such as Facebook and Twitter. The research tested different ML classifiers like Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) to categorize text from social media. The data used was posts from Facebook and Twitter. The research shows that SVM was the most accurate, reaching 78.99% accuracy. Similarly gathering a dataset from twitter, Khan [3]., conducted a case study on the official PlayStation account to illustrate their focus on sentiment analysis from a business intelligence standpoint, notably examining emotions and feelings expressed on Twitter. One thousand tweets from Twitter make up the dataset that was used. User emotions are compared between mentioning the official account and speaking generally as part of the analysis. When people contact their friends and followers instead of merely citing official accounts, the study finds that users are more open about their ideas and discontent. Also revolving on the idea of social media, Sánchez-Núñez et al. [35], provided an approach to model the business intelligence for identifying the users' abnormal emotion from their post through the social media marketing platform. To this end, to detect unusual manifestations in microblogs, the model uses the principles of multivariate Gaussian distribution and joint probability density. Data used in the study was obtained from micro-blogs, aggregated from 100 users over the period of 5 years with a sample size of 10,275. Thus, the accuracy rate that the presented model was achieving was approximately 83%. 87% for identification as an individual user or malicious bot. 84% for monthly identification of the subjects' abnormally endowed emotions.

In business intelligence and sentiment analysis, Swain & Cao [36], focused on utilizing sentiment analysis in order to extract the valuable insights from social media data related to supply chain management. The dataset which was used in the study includes and has over 600 randomly sampled companies from various industries, with data collected from forums, blogs, and microblogs such as Twitter. Tokenization, stemming, and feature selection utilizing filtering strategies such document frequency and mutual information are among the language processing approaches. With F-measures of 0.70 and 0.91 on the test set, respectively, the sentiment analysis approach uses a four-dimension classification algorithm and a positive-negative sentiment classification algorithm. Another study about sentimental analysis, Aqarwal [37]., aimed to have a better understanding about customer sentiments and goals in order to improve overall business strategy and overall customer satisfaction by applying DL models, namely Recurrent Neural Networks (RNNs) and a Convolutional Neural Networks (CNNs). The study does not specifically disclose the dataset that was utilized for analysis. English was the language used for sentiment analysis. The model uses deep learning techniques, such CNNs and RNNs, to effectively capture the subtleties of customer feedback, resulting in high accuracy rates in sentiment classification. Similarly focusing on customer satisfaction, Prananda & Thalib [38], focused on utilizing sentiment analysis and predictive analytics for business intelligence purposes, specifically in the context of analyzing customer reviews for GO-JEK services. It used predictive analytics and sentiment analysis for business intelligence, particularly when examining customer reviews for GO-JEK services. The dataset is made up of 3,111 tweets from various countries that had keywords associated with GO-JEK that were gathered in January 2019. For sentiment analysis classification, the research uses computer learning techniques like neural networks, SVMs, NB, and DT. The DT method does perform the best, it achieved a score of 0.55 in precision, recall, and f1-score. Almost Identically, Capuano [39], proposed a Hierarchical Attention Networks based approach for sentiment analysis application in customer relationship management. The model was trained on a dataset of more than 30,000 items, 40% of which were collected from an Italian IT company and 60% were collected from public datasets to balance the class distribution. The experimental results show that the proposed approach attains high accuracy rates, with a macro-averaged F1-score of 0.89 for the Italian language and 0.79 for English. The incremental learning mechanism does not affect the overall system performance, improving the model's performance over time.

In the same way, Srinivasan et al. [40], worked on the sentimental analysis of the impact of COVID-19 on social life using ML classifiers. The data which is used for the analysis is gathered from Twitter which contains manual sentiment labels on tweets like “Extremely Positive”, “Positive”, “Neutral”, “Negative” and “Extremely Negative”, it also has 41,158 rows of data. The data is collected from Kaggle COVID-19 NLP Text Classification dataset. They worked on the BERT model for the sentimental analysis and check the sentiment of the tweets from various countries and for India. Further focusing on the idea of NLP methods, Sanchez-Nunez et al. [41], concentrated on the opinion mining, sentiment analysis and emotion understanding particularly in the context of advertising. The research draws data from the WoS database and embraces articles published between 2010 and 2019. The study uses NLP and SA methods to analyze latent trends in consumer perceptions of international brands and products. Similarly, Gołębiowska et al. [42], elaborated on cybersecurity aspects in business intelligence analytics via sentiment analysis and big data. It addresses the need for dealing with large volumes of information and the corresponding needs and challenges of latest security demands. The analyzed dataset contains a broad range of information and may originate from different websites, social, scientific, political as well as business topics. The sentiment analysis should help to investigate user knowledge on information technologies, industry 4.0 and the related dangers and risks.

Similarly utilizing big data, Sreesurya et al. [43], centered on employing big data sentiment analysis with Hypex, an improved Long Short-Term Memory (LSTM) technique for retrieving business intelligence from reviews and comments. The dataset used to train is the 5-core Amazon dataset, which comprises of about 18 million reviews. The language processing method utilized entails converting text into word vectors using the GloVe model that has been trained previously. The model presents a testing accuracy rate as an indication of how the model performs when it comes to predicting the sentiment of the reviews. Hypex is the proposed activation function which surpasses general activation function for providing better accuracy in terms of sentiment classification for business intelligence. Further expanding on the idea of using big data, Niu et al. [44]., presented the Optimized Data Management using Big Data Analytics (ODM-BDA) model as a solution of optimizing organizational decision making by adopting big data technology and cloud computing strategies. It is intended to facilitate the Increase in efficiency in processing data, optimization of profits, and upgrade methods of decision. In this particular study, the improvement facilitated by the proposed ODM-BDA framework is achieved through a simulation analysis using 10 to 100 iterations to compare the enhancement in accuracy and duration.

Alike, Saura et al. [45]., employed a Latent Dirichlet Allocation (LDA) and Sentiment Analysis with SVM algorithm and to analyze the User Generated Content through Twitter for arriving at the critical factors that make a startup successful. The studied data set entails 35,401 tweets that include the #Startups hashtag. The analysis is performed through Python LDA 1. 0. 5 software and MonkeyLearn's Sentiment Analysis algorithm. In our model, the accuracy rate was obtained more than 0. 797 for positive sentiment, and >0. 802 for neutrality and above 0 for positive sentiment analysis. Similarly employing an SVM algorithm, Al-Otaibi et al. [46]., discussed about a system designed to measure customer satisfaction using sentiment analysis on Twitter data. The dataset used consists of 5513 hand-classified tweets, with positive and negative sentiments selected for training. The system employs the SVM classifiers for sentiment classification, achieving an accuracy rate of 87% on a 4000-tweet testing dataset.

In Arabic annotation, Guellil et al. [47], developed “ArAutoSenti” which aims to annotate Arabic text automatically based off its understanding on the sentiment behind the Arabic text. “ArAutoSenti” achieves an F1-score of 88%. It is important to mention that this method was applied for the under-resourced Algerian dialect. While Guelil [22], proposed SentiALG in this study, which is a tool intended for automatic annotation for the Algerian dialect in sentiment analysis. The dataset for it consisted of 8,000 messages. Correspondingly, Jarrar et al. [48], proposed an Arabic corpus called “SALMA”, which consists of around 34,000 tokens which are all sense-annotated. It is also worthy to mention that a smart web-based annotation tool was developed to support scoring multiple senses against a given word. In the same way, Al-Laith et al. [19]., presented a way to annotate large prices of texts in Arabic Corpus called “AraSenCorpus”. A neural network was used to train a set of models on a manually labeled dataset containing 15,000 tweets.

III. Methods and Materials

This section describes the methods that were used to carry out this study. There are six phases which were conducted namely; data collection, data cleaning, data annotation, data pre-processing, feature engineering, building models and model evaluation, all are as shown in Fig. 1.

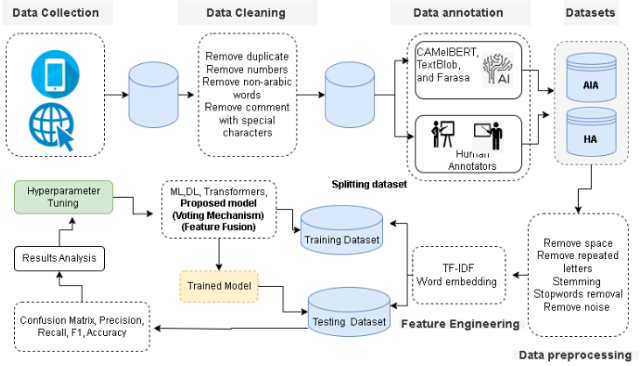

Fig. 1. Phases of the study.

A. Data Collection

In this phase, the data collected by user generated comments from the Play store of Mobile apps was through using the Python programming language. The main criteria are select the user comments and mobile apps are as follows: firstly, the period of collection for the user comments was between Jan 2022 to March 2024. Secondly, the mobile apps that are related to home delivery and transportation in total are eight Mobile apps. Lastly, the text for user comments is related to the Arabic language. All the user comments were downloaded into a separate CSV file for each application, then, all the files were combined together into one CSV file. At the end of this phase the total number of downloaded comments was approximately 223,341.

B. Data Cleaning

This phase had the task of preparing the dataset, in this phase several steps were used to remove duplicates of user comments, remove any user comments that are non- Arabic, remove the spaces from the files if there are any comments which are only empty, remove comments which consisted of special characters due to these types of comments being meaningless in the area of sentiment analysis.

C. Data Annotation

In the data annotation phase, there are two methods which have been utilized, namely; the first being the AIA tools which were used in the annotation process and the second being the HA method. In AIA, three tools have applied which are, TextBlob, Farasa, and CAMeLBERT. TextBlob is a python library built on Natural Language Toolkit (NLTK) and has been used for NLP tasks. Particularly “ar-textBlob” using the Standford API for Arabic techonizer. In this study it is used for sentiment analysis which indicates whether the user comments are positive or negative. Farasa at the Qatar Computing Research Institute has been developed for Arabic NLP. The CAMeLBERT is a pre-trained model based on the BERT architecture and specifically for Arabic NLP. In this type of annotation which is known as “ensemble techniques” is based on a voting mechanism which the decision is based on the majority. Therefore, if two classifiers indicated that the user comments belong to a specific class (either positive or negative) then the decision will be based on that. Thus, by the output of this method, the first dataset was constructed. Table I shows the description of the first dataset.

TABLE I. First Dataset Description (AIA)

Item(s)	No.comments	Min.Length	Max. Length
Positive	28,465	6	88
Negative	34,848	3	76
Total	63,313		

In the HA method, which has been conducted based on human annotation, three Arabic speakers helped in annotating the downloaded user comments. The user comments have been assigned to a class based on the decision of the majority, that means if two annotators agree about a specific decision on the user comments being assigned to the relevant class (either positive or negative) then the decision will be taken. Thus, the output of this method the second dataset has been constructed. Table II shows the description of the second dataset.

TABLE II. Dataset Description (HA)

Item(s)	No.comments	Min.Length	Max. Length
Positive	25,363	4	85
Negative	29,625	2	76
Total	54,988		

D. Feature Engineering

This phase is before feeding the models; each user comment is converted to numerical representation which is known as word representation. Such numerical representation is used as input to train and test the models. There are two types of word representation used in this study; the first being the Term-Frequency Inverse Document Frequency (TF-IDF) and word embedding’s. The TF-IDF is represent how the words is important in the dataset, it’s calculated as in mathematical Formula (1). While the second type of word embedding discovers the sematic relation between the words in user comments. Each word represented in vector dimension that contains values that measure the closeness of the word syntactic and semantic that happened in the training process.

$$TF-IDF(W,UC) = TF(W,UC)X\ IDF(W,UD) \quad (1)$$

Where, W is the word in the User Comment (UC). UD is represented in the dataset for all user comments. The TF is calculated based on the following mathematical Formula (2).

$$TF = \frac{F_{W,UC}}{N_{UC}} \quad (2)$$

Where $F_{W,UC}$, is the number of times the word (W) appears in the UC, and N represents the total number of W in UC.

While IDF is calculated as in the mathematical Formula (3).

$$IDF = \log \frac{N}{1+n_W} \qquad (3)$$

where, n_w is the number of UC that W appears in it.

E. Building Models

This subsection demonstrates the models that were used to evaluate the proposed models' performance compared. These models consisted of ML classifiers, DL models and transformers. The most common classifiers in ML were selected, which are NB, Logistics Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), DT and SVM [49]. While in DL experiments a LSTM, Bidirectional (BiLSTM), Gated Recurrent Unit (GRU), and CNN-LSTM were utilized, and in transform learning (transformers) [50], RoBERTa, MARBERT, distilbert, CAMeLBERT-DA, CAMeLBERT-Ca, AraELECTRA, ArabicBERT(Qarib), and AraBERTVer2 were utilized.

The proposed model consists of three transformers, which are built on the concept of transfer learning, as shown in Fig. 2. These models which were used in the proposed model are XLM-RoBERTa, AraBERT Ver2 and Arabic BERT. These models were trained on a large diverse Arabic dataset. These models are called pretrained models which can capture Arabic language patterns.

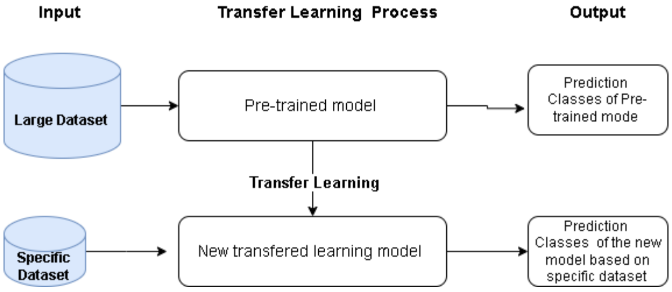
Input
Transfer Learning Process
Output
Large Dataset
Pre-trained model
Prediction Classes of Pre-trained mode
Transfer Learning
Specific Dataset
New transfered learning model
Prediction Classes of the new model based on specific dataset

Fig. 2. Transfer learning.

Additionally, these models are built on the BERT architecture. In this manner, the performance of the model improves. These pretrained models, called 'fine tuning models', can be utilised on a specific dataset. Therefore, these three state-of-the-art models can be utilised to benefit each one to improve the model's performance through SA on Arabic business intelligence. XLM-RoBERTa is a multilingual pretrained model which can be trained on a huge opera, including the Arabic language. It can help in discovering the linguistics between Arabic user comments and SA. By contrast, AraBERT Ver2 is an updated version of AraBERT and is essentially trained on Arabic text from different sources. It can handle and deeply understand the relation of Arabic words in the language's unique morphological characteristics. Arabic BERT is trained on a large dataset from different sources, and it is different from the datasets of AraBERT ver2.

The proposed model comprises of two methods, namely, the voting mechanism and feature fusion. In the voting mechanism, the models predict the class of the input comments, and each model gives its prediction separately. Then, the ensemble methods, i.e. the voting mechanism, are applied. In this manner, the decision is based on majority of the output of the pretrained models, as shown in Fig. 3.

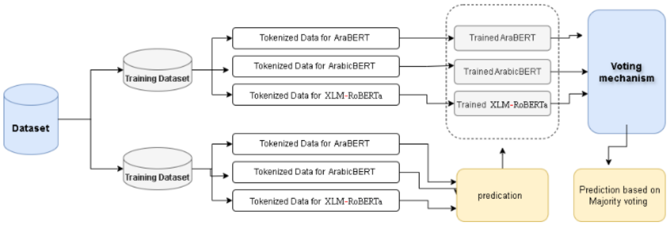
Dataset
Training Dataset
Training Dataset
Tokenized Data for AraBERT
Tokenized Data for ArabicBERT
Tokenized Data for XLM-RoBERTa
Tokenized Data for AraBERT
Tokenized Data for ArabicBERT
Tokenized Data for XLM-RoBERTa
Trained AraBERT
Trained ArabicBERT
Trained XLM-RoBERTa
Voting mechanism
predication
Prediction based on Majority voting

Fig. 3. Voting mechanism method.

The second method is feature fusion, which works by receiving the vectors from three pretrained models. Then, these vectors are combined to a large vector that is subsequently used to feed fully connected layers that are produced in the final prediction of the classes. Fig. 4 illustrates the feature fusion method.

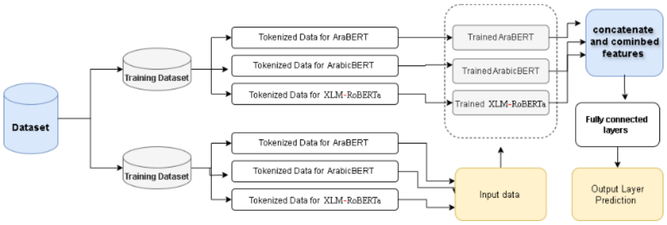
Dataset
Training Dataset
Training Dataset
Tokenized Data for AraBERT
Tokenized Data for ArabicBERT
Tokenized Data for XLM-RoBERTa
Tokenized Data for AraBERT
Tokenized Data for ArabicBERT
Tokenized Data for XLM-RoBERTa
Trained AraBERT
Trained ArabicBERT
Trained XLM-RoBERTa
concatenate and cominbed features
Fully connected layers
Input data
Output Layer Prediction

Fig. 4. Feature fusion method.

F. Model Performance Evaluation

In order to validate the proposed method of automatic Arabic annotation and also the proposed model of hybrid transformers models (transfer learning). The models' common measurement matrix has been used which is the precision, recall, F1, Accuracy and AUC-ROC.

Precision is the ratio of correctly predicted positive observations to the total predicted positives. Precision is calculated based on the correctly positive user comments (UCs) predicted to the total number of predicted positive UCs as presented in mathematical Formula (4). In the instance in which the precision is high, that means the model predicted the positive or negative UCs correctly. Recall is the ratio of the True positive UCs predicted to the actual positive UCs as displayed in the mathematical Formula (5). The F1-score or also called the F-measure is the homogenous between the precision and recall. While the accuracy is the total number of correct predicted positive UCs over the total number of UCs in the dataset as portrayed in the mathematical Formula (7).

$$\text{Precision} = \frac{True\ positive\ UCs}{True\ positive\ UCs + false\ positive\ UCs} \qquad (4)$$

$$\text{Recall} = \frac{True\ positive\ UCs}{True\ positive\ UCs + false\ negative\ UCs} \qquad (5)$$

$$\text{F1} = 2X\ \frac{PrecisionXRecall}{Precsiosn + Recall} \qquad (6)$$

$$\text{Accuracy} = \frac{True\ positive\ UCs+True\ negative\ UCs}{Total\ number\ of\ UCs\ is\ dataset} \tag{7}$$

AUC-ROC is the graphical representation that shows the model performance in distinguishing between the positive user comments and the negative UCs. AUC-ROC is calculated based on the True Positive Rate (TPR) as presented in mathematical Formula (8) and False Positive Rate (FPR) as shown in mathematical Formula (9). The final value is between 0 and 1, the higher the values are, indicates how well the model has performed.

$$TPR = \frac{True\ postives\ UCs\ for\ both\ classes}{True\ Psotive+Flase\ Negative} \tag{8}$$

$$FPR = \frac{False\ \ postives\ UCs\ for\ both\ classes}{False\ \ psotive+True\ Negative} \tag{9}$$

IV. Results and Experiments

This section describes the settings of the experiments for the four types of experiments conducted: ML experiments, DL experiments, transformer experiments and the proposed model experiments. The experiment results are then explained.

A. *Experimental Settings*

All the experiments are conducted using Google Colab, utilising the GPU and other hardware-related matters. The programming language of choice is Python. In the ML experiment, the scikit-learn package is used to split the dataset and to import and use the ML classifiers. In the DL experiments, the TensorFlow framework is used for building DL models. The transformer package is imported to conduct the transformers experiments, utilising the hugging face platform to access the pretrained models. In all the experiments, the dataset is divided into two parts, 70% for training and 30% for testing the models. Hyperparameters for ML, DL, transformers and the proposed model is presented in Tables III, IV and V respectively.

TABLE III. ML Hyperparameters

Classifier	Values
DT	Criterion’gini'
KNN	n_neighbors 5
LR	penalty 'l2'
RF	n_estimators 100
SVM	C 1.0

TABLE IV. DL Hyperparameters

Parameter	Value
LSTM Units	64
Dropout	0.5
Batch Size	32
Optimizer	Adam
Activation function (Hidden Layers)	ReLU
Activation function (output)	Sigmod

TABLE V. Hyperparameters Transformers

Item(s)	Values
Batch Size	16
Number of Epochs	6
weight_decay	0.01
logging_steps	100
learning_rate	2e-5

B. *Experimental Results*

In this subsection, the results of the four experiments are analysed. In the ML experiments, several experiments are conducted using the AIA and HA methods. Table VI shows the results based on the most commonly used measurements, i.e. precision, recall, F1-score and accuracy.

TABLE VI. Comparison Between the Four Measurements Using the AIA Method

Classifiers	Precision	Recall	F1-Score	Accuracy
DT	81.53%	64.19%	65.64%	78.05%
KNN	73.45%	75.90%	74.28%	77.41%
LR	83.08%	80.27%	81.46%	85.21%
NB	74.91%	77.88%	75.84%	78.59%
RF	81.99%	80.76%	81.33%	84.77%
SVM	82.82%	79.77%	81.04%	84.93%

Table VI presents the performance metrics of various classifiers used for SA in the context of business intelligence, specifically in extracting insights from user comments on mobile apps. LR shows the highest accuracy of 85.21% and an F1-score of 81.46%, indicating its strong ability to classify sentiments from user comments accurately. RF and SVM also perform well, with accuracies of 84.77% and 84.93%, respectively, and F1-scores exceeding 81%. By contrast, the DT classifier has the lowest recall at 64.19% and F1-score at 65.64%, suggesting that it may not be as effective for this application compared with the other models. KNN and NB show moderate performance, with KNN achieving a recall of 75.90% and an F1-score of 74.28%, whereas NB shows a balanced performance with a precision of 74.91% and an F1-score of 75.84%. The confusion matrix is shown in Fig. 5, and the AUC–ROC is presented in Fig. 6. When the HA method is used, the best accuracy is recorded when using the LR classifier and low accuracy is achieved using KNN. Table VII shows the comparison between the four measurements using the HA method for the ML classifiers. The confusion matrix and AUC–ROC are presented in Fig. 7 and Fig. 8, respectively.

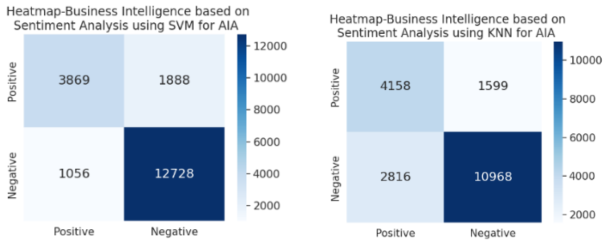

Fig. 5. Confusion matrix SVM and KNN using AIA method.

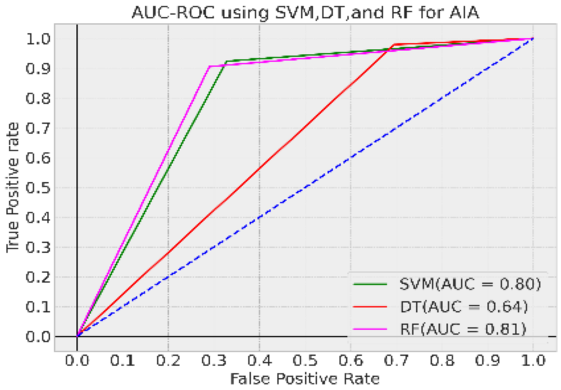

Fig. 6. AUC-ROC- AIA method.

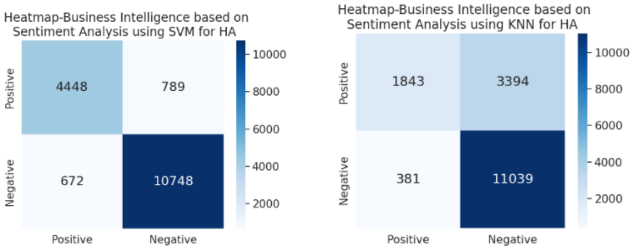

Fig. 7. Confusion matrix SVM and KNN using HA method.

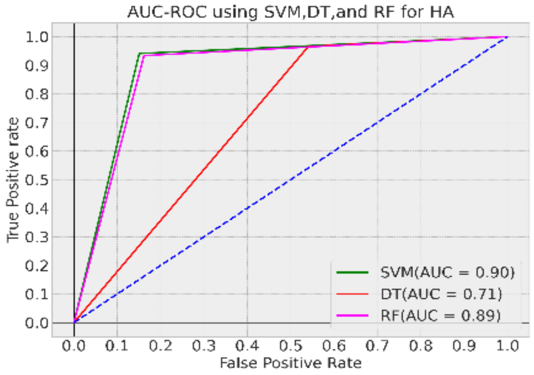

Fig. 8. AUC-ROC –HA method.

TABLE VII. THE COMPARISON BETWEEN THE FOUR MEASUREMENTS USING HA METHOD

Classifiers	Precision	Recall	F1-Score	Accuracy
DT	82.75%	71.25%	73.56%	80.64%
KNN	79.68%	65.93%	67.40%	77.34%
LR	89.96%	89.13%	89.53%	91.06%
NB	83.00%	86.13%	84.07%	85.54%
RF	88.99%	88.61%	88.80%	90.39%
SVM	90.02%	89.52%	89.76%	91.23%

Table VII presents the performance metrics of various classifiers used for SA in the context of business intelligence, focusing on user comments from mobile apps. LR and SVM stand out with the highest performance, achieving accuracy rates of 91.06% and 91.23%, respectively, and F1-scores of 89.53% and 89.76%. R also shows strong performance with an accuracy of 90.39% and an F1-score of 88.80%. NB demonstrates balanced performance with an accuracy of 85.54% and an F1-score of 84.07%. By contrast, DT and KNN perform less effectively, with DT achieving an accuracy of 80.64% and an F1-score of 73.56%, whilst KNN has the lowest accuracy at 77.34% and an F1-score of 67.40%. Overall, LR, SVM, and RF are the most effective classifiers for extracting sentiment-based insights from user comments, highlighting their suitability for enhancing business intelligence through SA.

In the DL experiments, specifically during the utilisation of the AIA method, GRU demonstrates the highest performance with an accuracy of 90.01% and an F1-score of 87.93%, indicating its effectiveness in accurately classifying sentiments from user comments. The accuracy of training and validation is presented in Fig. 9. fNN-LSTM, whilst still effective, shows a slightly lower performance with an accuracy of 89.37% and an F1-score of 87.34%.

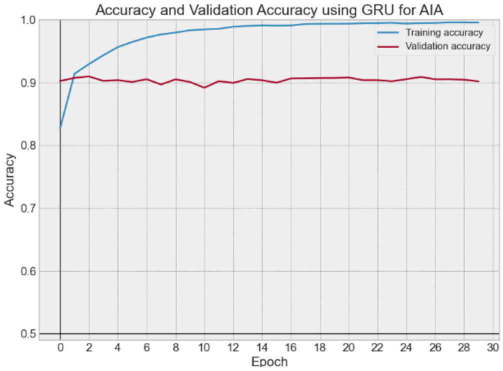

Fig. 9. Accuracy of training and validation using GRU for AIA.

In the HA method, LSTM demonstrates the highest performance with an accuracy of 95.54% and an F1-score of 94.85%, making it the most effective in accurately classifying sentiments from user comments. The accuracy of training and validation is shown in Fig. 10. GRU follows closely with an accuracy of 95.38% and an F1-score of 94.64%, indicating its strong potential as well. CNN-LSTM, whilst showing slightly lower recall, maintains high precision and F1-score with an accuracy of 95.09%. BiLSTM, although slightly behind the others, still performs robustly with an accuracy of 94.94% and consistent precision, recall and F1-score of 94.14%. Fig. 11 shows the precision, recall, F1-score and accuracy using the AIA and HA methods.

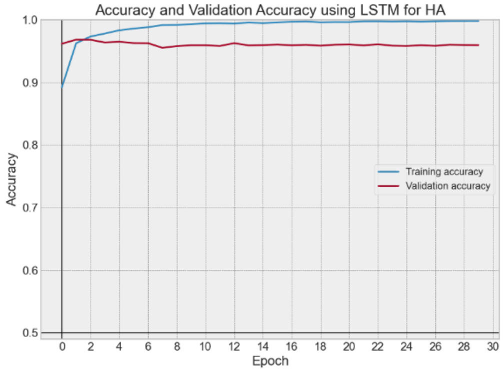

Fig. 10. Accuracy of training and validation using LSTM for HA.

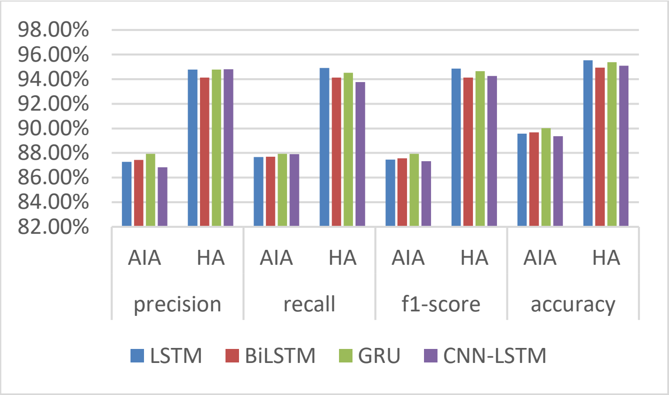

Fig. 11. Model performance between the AIA and HA methods.

In the pretrained models (transformers) and the proposed models, the experiments were conducted using the AIA and HA methods. The experimental results of the AIA and HA methods for the pretrained (transformers) and the proposed models is presented in Table VIII. In the proposed model, two methods, namely, voting mechanism and feature fusion, are used.

Table VIII summarises the performance of the pretrained models (transformers) used for the SA of user comments and reviews on mobile apps for enhancing business intelligence. Each model was evaluated across two approaches: AIA and HA. Amongst the models, ArabBERTVEr2 achieved an accuracy of 92.48% in AIA and 97.06% in AIA. The proposed Model 1 (feature fusion) demonstrated the highest accuracy with 93.96% AIA and 98.11% HA. These accuracies reflect how effectively each model can classify sentiments expressed in user feedback, providing valuable insights for improving mobile app performance and user satisfaction in business contexts. Additionally, the proposed Model 2 (voting mechanism) achieved an accuracy of 92.65% and 97.24%. The confusion matrix and AUC–ROC for the proposed model using the feature fusion method is presented in Fig. 12 and Fig. 13, respectively. While Fig. 14 and Fig. 15 shows confusion matrix and AUC-ROC of the proposed model using the voting mechanism respectively.

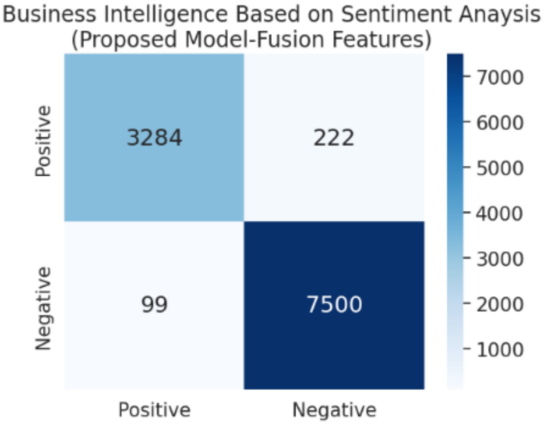

Fig. 12. Confusion matrix of the proposed model using feature fusion.

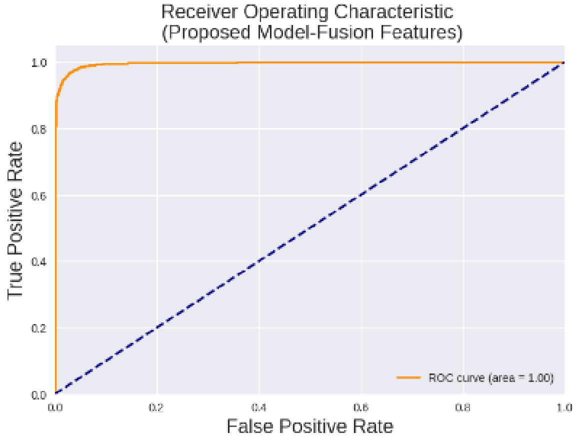

Fig. 13. AUC–ROC of the proposed model using feature fusion.

TABLE VIII. COMPARISON BETWEEN THE PRECISION, RECALL, F1-SCORE AND ACCURACY FOR TRANSFORMERS AND THE PROPOSED MODEL

Model(s)	**Precision**		**Recall**		**F1-score**		**Accuracy**	
	AIA	**HA**	**AIA**	**HA**	**AIA**	**HA**	**AIA**	**HA**
ArabBERTVEr2	93.43%	97.50%	96.13%	98.22%	94.76%	97.86%	92.48%	97.06%
ArabicBERT-Qarib	90.37%	97.01%	95.46%	98.26%	92.84%	97.63%	89.58%	96.74%
AraELECTRA	92.23%	97.34%	96.08%	98.34%	94.12%	97.84%	91.50%	97.03%
CAMeLBERT-DA	94.40%	96.86%	94.99%	97.47%	94.69%	97.17%	92.46%	96.11%
CAMeLBERT-Ca	92.57%	97.05%	95.59%	97.78%	94.06%	97.41%	91.45%	96.44%
distilbert	93.75%	97.57%	96.05%	98.22%	94.89%	97.89%	92.68%	97.11%
MARBERT	94.10%	97.27%	95.39%	97.59%	94.74%	97.43%	92.51%	96.48%
MERTICRobBERt	90.78%	97.33%	97.09%	98.20%	93.83%	97.76%	90.96%	96.92%
Proposed Mode 1	93.94%	98.22%	92.78%	97.94%	93.36%	98.07%	**93.96%**	**98.11%**
Proposed Mode 2	92.24%	97.62%	93.12%	96.95%	92.66%	97.28%	**92.65%**	**97.24%**

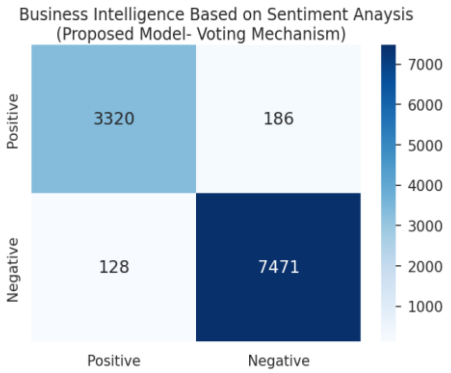

Fig. 14. Confusion matrix of the proposed model using the voting mechanism.

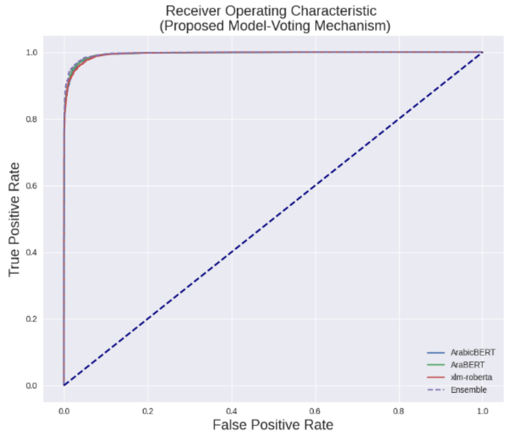

Fig. 15. AUC–ROC of the proposed model using the voting mechanism.

V. Discussion

This study has two findings related to the proposed hybrid transform learning model and the proposed automatic Arabic annotation methods.

Firstly, the experiment results reveal that the performance of the proposed model outperforms the aforementioned experiments, i.e. DL, ML, and transformers experiments. To go into depth of the precise method in which the proposed model achieved the highest results in, is the feature fusion which attained an accuracy score of 98.11%, which proves that it is better equipped for identifying the sentiment behind user comments in the Arabic language on whether the sentiment behind the written comment is positive or negative than the voting mechanism which reached an accuracy if 97.24% respectively. This study demonstrates that utilising transfer learning (pretrained models/transformers) enhances the model's performance in terms of accuracy in identifying and distinguishing positive and negative sentiments in Arabic user comments. The reason is that the pretrained models have been trained on large datasets that help in exploring and discovering the semantic and syntactic relationships between Arabic words.

Secondly, in the proposed Arabic automatic method, AIA saves time and effort in annotating Arabic user comments because it is completely automated as opposed to the HA method, which requires much more effort in annotating Arabic user comments and at least three native Arabic speakers to perform the voting mechanism. Nevertheless, the HA method achieves a somewhat higher accuracy than the AIA method, whilst the difference between the two methods in performance is approximately 5%. This result is observed in all the conducted experiments.

VI. Conclusion

Transformers have been utilised to improve the capabilities in handling natural language processing. Therefore, this study proposes a robust hybrid transfer learning model to enhance business intelligence by accurately detecting users' sentiments. The model combines XLM-RoBERTa, AraBERT Ver2 and Arabic BERT, and the model additionally utilises a voting mechanism and feature fusion to improve the models' performance. Additionally, the study introduces AIA, which integrates CAMeLBERT, TextBlob and Farasa. Furthermore, a novel dataset of user-generated comments from mobile apps is introduced. Results demonstrate that the proposed model leveraging feature fusion and the voting mechanism outperforms all baselines with an accuracy of 97.24% and 98.11%, respectively. Furthermore, the AIA method is closely matched with the HA in the Arabic corpora annotation, confirming its reliability and accuracy. In the future work, enhanced Arabic annotation methods are to be through utilising transformers and to build lexical dictionary, expanding the dataset to include more diverse sources and languages that could help generalise the model's applicability across various domains. Apply the automatic annotation to Arabic dialects across the 22 Arabic-speaking countries, each Arabic speaking country with its own unique variations. This approach can significantly enhance the classification process.

References

[1] Kurnia, P. F. (2018). "Business intelligence model to analyze social media information". Procedia Computer Science, 135, 5-14.

[2] Mehta, P., & Pandya, S. (2020). "A review on sentiment analysis methodologies, practices and applications". International Journal of Sci entific and Technology Research, 9(2), 601-609.

[3] Khan, S. (2022, February). "Business Intelligence Aspect for Emotions and Sentiments Analysis". In 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) (pp. 1-5). IEEE.

[4] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). "A survey on sentiment analysis methods, applications, and challenges". Artificial Intelligence Review, 55(7), 5731-5780.

[5] Gohil, S., Vuik, S., & Darzi, A. (2018). "Sentiment analysis of health care tweets: review of the methods used". JMIR public health and surveillance, 4(2), e5789.

[6] Khan, M. T., & Khalid, S. (2016). "Sentiment analysis for health care. In Big data: concepts, methodologies, tools, and applications", (pp. 676-689). IGI Global.

[7] Aattouchi, I., Elmendili, S., & Elmendili, F. (2021). "Sentiment Analysis of Health Care". In E3S Web of Conferences (Vol. 319, p. 01064). EDP Sciences.

[8] Hajrizi, R., & Nuçi, K. P. (2020). "Aspect-based sentiment analysis in education domain". arXiv preprint arXiv:2010.01429.

[9] Shanthi, I. (2022). "Role of educational data mining in student learning processes with sentiment analysis: A survey". In Research Anthology on Interventions in Student Behavior and Misconduct (pp. 412-427). IGI Global.

[10] Alhujaili, R. F., & Yafooz, W. M. (2022, May). "Sentiment analysis for youtube educational videos using machine and deep learning

approaches". In 2022 IEEE 2nd international conference on electronic technology, communication and information (ICETCI) (pp. 238-244). IEEE.

[11] Lin, H. C. K., Wang, T. H., Lin, G. C., Cheng, S. C., Chen, H. R., & Huang, Y. M. (2020). "Applying sentiment analysis to automatically classify consumer comments concerning marketing 4Cs aspects". Applied Soft Computing, 97, 106755.

[12] Reyes-Menendez, A., Saura, J. R., & Filipe, F. (2020). "Marketing challenges in the# MeToo era: Gaining business insights using an exploratory sentiment analysis". Heliyon, 6(3).

[13] Mehraliyev, F., Chan, I. C. C., & Kirilenko, A. P. (2022). "Sentiment analysis in hospitality and tourism: a thematic and methodological review". International Journal of Contemporary Hospitality Management, 34(1), 46-77.X.4

[14] Ahmed, A. A. A., Agarwal, S., Kurniawan, I. G. A., Anantadjaya, S. P., & Krishnan, C. (2022). "Business boosting through sentiment analysis using Artificial Intelligence approach". International Journal of System Assurance Engineering and Management, 13(Suppl 1), 699-709.

[15] Sudirjo, F., Diantoro, K., Al-Gasawneh, J. A., Azzaakiyyah, H. K., & Ausat, A. M. A. (2023). "Application of ChatGPT in Improving Customer Sentiment Analysis for Businesses". Jurnal Teknologi Dan Sistem Informasi Bisnis, 5(3), 283-288.

[16] Yin, J. Y. B., Saad, N. H. M., & Yaacob, Z. (2022). "Exploring Sentiment Analysis on E-Commerce Business: Lazada and Shopee". Tem journal, 11(4), 1508-1519.

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

[18] Renault, T. (2020). "Sentiment analysis and machine learning in finance: a comparison of methods and models on one million messages". Digital Finance, 2(1), 1-13.

[19] Al-Laith, A., Shahbaz, M., Alaskar, H. F., & Rehmat, A. (2021). "Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus". Applied Sciences, 11(5), 2434.

[20] Almuzaini, H. A., & Azmi, A. M. (2022). "An unsupervised annotation of Arabic texts using multi-label topic modeling and genetic algorithm". Expert Systems with Applications, 203.

[21] Almuqren, L., Alzammam, A., Alotaibi, S., Cristea, A., & Alhumoud, S. (2017). "A review on corpus annotation for Arabic sentiment analysis. In Social Computing and Social Media. Applications and Analytics" 9th International Conference, SCSM 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II 9 (pp. 215-225). Springer International Publishing. 117384.

[22] Guellil, I., Adeel, A., Azouaou, F., & Hussain, A. (2018). "Sentialg: Automated corpus annotation for algerian sentiment analysis. In Advances in Brain Inspired Cognitive Systems", 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings 9 (pp. 557-567). Springer International Publishing.

[23] Al-Laith, Ali, Muhammad Shahbaz, Hind F. Alaskar, and Asim Rehmat. "Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus." Applied Sciences 11, no. 5 (2021): 2434.

[24] Habash, N., & Palfreyman, D. (2022, June). "ZAEBUC: An annotated Arabic-English bilingual writer corpus". In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 79-88).

[25] Alahmary, R. M., Al-Dossari, H. Z., & Emam, A. Z. (2019, January). "Sentiment analysis of Saudi dialect using deep learning techniques", In 2019 International Conference on Electronics, Information, and Communication (ICEIC) (pp. 1-6). IEEE.

[26] Rahab, H., Zitouni, A., & Djoudi, M. (2021). "SANA: Sentiment analysis on newspapers comments in Algeria". Journal of King Saud University-Computer and Information Sciences, 33(7), 899-907.

[27] Al-Thubaity, A., Alharbi, M., Alqahtani, S., & Aljandal, A. (2018, April). "A Saudi dialect Twitter Corpus for sentiment and emotion analysis". In 2018 21st Saudi computer society national computer conference (NCC) (pp. 1-6). IEEE.

[28] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2020). "ASA: A framework for Arabic sentiment analysis". Journal of Information Science, 46(4), 544-559.

[29] Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., & Al Kaabi, M. (2018, May). "A morphologically annotated corpus of Emirati Arabic", In Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).

[30] Elnagar, A., & Einea, O. (2016, November). "BRAD 1.0: Book reviews in Arabic dataset". In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-8). IEEE.

[31] Elnagar, A., Lulu, L., & Einea, O. (2018). "An annotated huge dataset for standard and colloquial arabic reviews for subjective sentiment analysis". Procedia computer science, 142, 182-189.

[32] Gamal, D., Alfonse, M., El-Horbaty, E. S. M., & Salem, A. B. M. (2019). "Twitter benchmark dataset for Arabic sentiment analysis". Int J Mod Educ Comput Sci, 11(1), 33.

[33] Abdellaoui, H., & Zrigui, M. (2018). "Using tweets and emojis to build tead: an Arabic dataset for sentiment analysis". Computación y Sistemas, 22(3), 777-786.

[34] Abo, M. E. M., Shah, N. A. K., Balakrishnan, V., Kamal, M., Abdelaziz, A., & Haruna, K. (2019, April). "Ssa-sda: subjectivity and sentiment analysis of sudanese dialect Arabic". In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-5). IEEE.

[35] Sánchez-Núñez, P., Cobo, M. J., De Las Heras-Pedrosa, C., Peláez, J. I., & Herrera-Viedma, E. (2020). "Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis". IEEE Access, 8, 134563-134576.

[36] Swain, A. K., & Cao, R. Q. (2019). "Using sentiment analysis to improve supply chain intelligence". Information Systems Frontiers, 21, 469-484.

[37] Agarwal, S. (2022). "Deep learning-based sentiment analysis: Establishing customer dimension as the lifeblood of business management". Global Business Review, 23(1), 119-136.

[38] Prananda, A. R., & Thalib, I. (2020). "Sentiment analysis for customer review: Case study of GO-JEK expansion". Journal of Information Systems Engineering and Business Intelligence, 6(1), 1.

[39] Capuano, N., Greco, L., Ritrovato, P., & Vento, M. (2021). "Sentiment analysis for customer relationship management: an incremental learning approach". Applied intelligence, 51, 3339-3352.

[40] Srinivasan, S. M., Shah, P., & Surendra, S. S. (2021). "An approach to enhance business intelligence and operations by sentimental analysis". Journal of System and Management Sciences, 11(3), 27-40.

[41] Sánchez-Núñez, P., Cobo, M. J., De Las Heras-Pedrosa, C., Peláez, J. I., & Herrera-Viedma, E. (2020). "Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis". IEEE Access, 8, 134563-134576.

[42] Gołębiowska, A., Jakubczak, W., Prokopowicz, D., & Jakubczak, R. (2021). "Cybersecurity of business intelligence analytics based on the processing of large sets of information with the use of sentiment analysis and Big Data". European Research Studies Journal, 24(4).

[43] Sreesurya, I., Rathi, H., Jain, P., & Jain, T. K. (2020). "Hypex: A tool for extracting business intelligence from sentiment analysis using enhanced LSTM". Multimedia Tools and Applications, 79, 35641-35663.

[44] Niu, Y., Ying, L., Yang, J., Bao, M., & Sivaparthipan, C. B. (2021). "Organizational business intelligence and decision making using big data analytics". Information Processing & Management, 58(6), 102725.

[45] Saura, J.R.; Palos-Sanchez, P.; Grilo, A. Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining. Sustainability 2019, 11, 917.

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

[47] Guellil, I., Azouaou, F., & Chiclana, F. (2020)." ArAutoSenti: automatic annotation and new tendencies for sentiment classification of Arabic messages". Social Network Analysis and Mining, 10, 1-20.

[48] Jarrar, M., Malaysha, S., Hammouda, T., & Khalilia, M. (2023). "Salma: Arabic sense-annotated corpus and wsd benchmarks". arXiv preprint arXiv:2310.19029.

[49] Alhejaili, R., Alhazmi, E. S., Alsaeedi, A., & Yafooz, W. M. (2021, September). “Sentiment analysis of the COVID-19 vaccine for Arabic tweets using machine learning”. In 2021 9th International conference on reliability, infocom technologies and optimization (Trends and Future Directions)(ICRITO) (pp. 1-5). IEEE.

[50] Yafooz, W. M., Al-Dhaqm, A., & Alsaeedi, A. (2023). “Detecting kids cyberbullying using transfer learning approach: Transformer fine-tuning models”. In Kids Cybersecurity Using Computational Intelligence Techniques (pp. 255-267). Cham: Springer International Publishing.