Social Network Analysis and Mining (2023) 13:38
https://doi.org/10.1007/s13278-023-01043-6

ORIGINAL ARTICLE

An enhanced approach for sentiment analysis based on meta-ensemble deep learning

Rania Kora[1] · Ammar Mohammed[1]

Received: 31 December 2022 / Revised: 27 January 2023 / Accepted: 17 February 2023

Abstract
Sentiment analysis, commonly known as "opinion mining," aims to identify sentiment polarities in opinion texts. Recent years have seen a significant increase in the acceptance of sentiment analysis by academics, businesses, governments, and several other organizations. Numerous deep-learning efforts have been developed to effectively handle more challenging sentiment analysis problems. However, the main difficulty with deep learning approaches is that they require a lot of experience and hard work to tune the optimal hyperparameters, making it a tedious and time-consuming task. Several recent research efforts have attempted to solve this difficulty by combining the power of ensemble learning and deep learning. Many of these efforts have concentrated on simple ensemble techniques, which have some drawbacks. Therefore, this paper makes the following contributions: First, we propose a meta-ensemble deep learning approach to improve the performance of sentiment analysis. In this approach, we train and fuse baseline deep learning models using three levels of meta-learners. Second, we propose the benchmark dataset "Arabic-Egyptian Corpus 2" as an extension of a previous corpus. The corpus size has been increased by 10,000 annotated tweets written in colloquial Arabic on various topics. Third, we conduct several experiments on six benchmark datasets of sentiment analysis in different languages and dialects to evaluate the performance of the proposed meta-ensemble deep learning approach. The experimental results reveal that the meta-ensemble approach effectively outperforms the baseline deep learning models. Also, the experiments reveal that meta-learning improves performance further when the probability class distributions are used to train the meta-learners.

Keywords Ensemble learning · Ensemble deep learning · Ensemble methods · Deep learning · Sentiment analysis

1 Introduction

The power of social media for expressing opinions about events, topics, people, services, or products has expanded due to the growth of user-generated content on platforms (Naresh and Venkata Krishna 2021). Hence, analyzing this huge amount of social media data can help better understand public opinions and trends and effectively make important decisions by classifying the opinions and feelings expressed in the text and determining their polarity as positive, negative, or neutral (Mejova 2009).

In the literature, several research efforts have been introduced to approach sentiment analysis using machine learning (Pontiki et al. 2016; Ahmed et al. 2013; Duwairi et al. 2014; Shoukry and Rafea 2012; Alomari et al. 2017). Extended efforts have used deep learning to handle bigger data and improve the classification's performance against classical machine learning models (Mohammed and Kora 2019; Chen et al. 2018; Pontiki et al. 2016; Heikal et al. 2018; Baly et al. 2017; Rojas-Barahona 2016). Deep learning techniques aim to overcome the limitations and problems of classical learning through efficient approaches in dealing with complex problems, large amounts of data, and its capacity to automatically extract the feature from the text (Habimana et al. 2020; Chan et al. 2020). There are several architectures and models for deep learning approaches when applied to sentiment analysis, such as recurrent neural networks (RNN) (Moitra and Mandal 2019), gated recurrent unit (GRU) (Le et al. 2019), Long Short-Term Memory (LSTM) (Graves

✉ Ammar Mohammed
ammar@cu.edu.eg

Rania Kora
rania.kora@pg.cu.edu.eg

[1] Department of Computer Science, Faculty of Graduate Studies for Statistical Researches, Cairo University, Cairo, Egypt

Published online: 02 March 2023

2012), Convolutional Neural Networks (CNN) (Collobert and Weston 2008). However, the main difficulty with deep learning techniques is identifying the most appropriate architectures and models. Usually, deep models require much effort due to tuning the optimal hyperparameters in the search space of the possible hyperparameters, which is a tedious task (Yadav and Vishwakarma 2020). These problems can be overcome by approaching ensemble learning to deep learning. Traditional ensemble learning refers to merging several basic models to build one powerful model (Kumar et al. 2021). Ensemble learning has been successfully applied in many fields, such as image classification (Wang et al. 2013), medical image (Cho and Won 2003; Shipp and Kuncheva 2002), music recognition (Stamatatos and Widmer 2002), malware detection (Shahzad and Lavesson 2013) and text classification (Kulkarni et al. 2018). In the literature, there are several ensemble approaches, like, averaging, boosting, bagging, random forest, and stacking (Zhang and Ma 2012). In deep learning, most ensemble learning is a simple averaging of model (Tan et al. 2022; Mohammadi and Shaverizade 2021; Araque et al. 2017) due to its simplicity and high results. However, the voting-based ensemble method is not a smart method to combine the models because it is biased toward weak models, which can reduce the performance in a lot of problems (Tasci et al. 2021).

To this end, the primary objectives of this research are four-fold. First, we propose a meta-ensemble deep learning approach to boost the performance of sentiment analysis. The proposed approach combines the predictions of several groups of deep models using three levels of meta-learners. In the proposed approach, we achieve diversity in the ensemble by using differences in the training data, the diversity of trained baseline deep learners, and the variation within the fusion of baseline deep models. Second, we propose the benchmark dataset "Arabic-Egyptian corpus", which consists of 50,000 tweets written in colloquial Arabic on various topics. This corpus is an extended version of the corpus "Arabic-Egyptian corpus" (Mohammed and Kora 2019). Third, we conduct a wide range of experiments on six public benchmark datasets to study the performance of the proposed meta-ensemble deep learning approach on sentiment classification in different languages and dialects. For each benchmark dataset, groups of different deep baseline models are trained on partitions of the trained data. Their best performance is compared with the proposed meta-ensemble deep learning approach. Finally, we show the impact of meta-predictions of the proposed meta-ensemble deep learning approach through different models' predictions, namely the class label probability distribution and the class label predictions. The main contributions of the paper can be summarized as follows:

- We propose a meta-ensemble deep learning approach to improve the sentiment classification performance that combines three levels of meta-learners.
- We extended the Arabic-Egyptian corpus (Mohammed and Kora 2019) by increasing it to 50k annotated tweets.
- We train several baseline deep models using six public benchmark sentiment analysis datasets in different languages and dialects.
- We conduct a wide range of experiments to study the effect of the meta-ensemble deep learning approach against single deep learning models.
- We compare the effect of the generated predictions of meta-learners involved in the proposed approach to improve the performance.

The paper is structured as follows: Sect. 2 provides a brief overview of the challenges of sentiment analysis and various ensemble learning methods as well as highlighting some of the literature used for ensemble learning in sentiment analysis. Section 3 describes the meta-ensemble deep learning approach. Section 4 shows the experimental results, the evaluation of the baseline deep learning models, and the meta-ensemble deep learning approach in each of the different benchmark datasets. Finally, Sect. 5 concludes the paper and suggests future research directions.

2 Related work

Through sentiment analysis, we can obtain important information that helps in making decisions, solving problems, managing crises, correcting misconceptions, providing desired products and services, interacting with consumers on their terms, improving product and service quality, discovering new marketing strategies and increasing sales (Tuysuzoglu et al. 2018). Despite its benefits, sentiment analysis is an extremely difficult task due to several challenges and problems (Cambria et al. 2017). First, the problem of identifying the subjective parts of the text: The same word can be treated as subjective in one context, while it might be objective in some other. This makes it challenging to distinguish between subjective and objective (sentiment-free) texts. For instance: "The writer's language was very crude," and "Crude oil is extracted from the sea-beds". Second, the problem of domain Dependence: In other contexts, the same sentence can indicate something quite different. The word unpredictable is negative in the domain of movies, but when used in another context, it has a positive connotation. For instance: "The movie was too slow and too long", "I love long pasta". Third, the problem of sarcasm Detection: Sarcastic sentences use positive words to convey a negative opinion about a target. For instance: "Nice perfume. You must be marinated in it". Fourth, the problem of thwarted

Expressions: In some sentences, the polarity of the text is determined by a small portion of the text. For instance: “Although I’m tired, the day is great.” Fifth, the problem of indirect Negation of Sentiment: Such negations are not easily defined because they do not contain “no,” “not,” etc. Sixth, the problem of order Dependence: When the words are not considered independent. For instance, “A is better than B”. Seventh, the problem of entity Recognition: A text may not always refer to the same entity. For instance, “I hate Samsung, but I like OPPO”. Eighth, the problem of identifying Opinion Holders: All written in a text is not always the author’s opinion. For instance, when the author quotes someone. Ninth and finally, the problem of associating sentiment with specific keywords: Many statements express very strong opinions, but it is impossible to identify the source of these sentiments. Generally, sentiment analysis can occur at three levels: Sentence, Document, and Aspect/Feature. At the sentence level, the task of this level is sentence by sentence and decides whether each sentence represents a neutral, positive, or negative opinion. At the document level, this analysis level identifies a document’s overall sentiment and categorizes it as negative or positive. At the aspect level (also known as a word or feature level), this level of analysis aims to discover sentiments on entities and/or their aspects (Wagh and Punde 2018).

In recent years, ensemble learning has been considered one of the most successful techniques in machine learning (Sagi and Rokach 2018). The main factors behind the ensemble system’s success are increasing diversity among baseline classifier types, using different ensemble methods, using different beginning parameters, and creating multiple datasets from the original dataset (cross-validation or sub-samples) (Mohammed and Kora 2021). Ensemble methods aim to increase prediction accuracy by combining decisions from various sub-models into a new model. Besides, the ensemble methods help avoid overfitting and reduce variance and biases. Also, ensemble learning helps to generate multiple hypotheses using the same base learner. In addition, ensemble learning methods help reduce the drawbacks of the baseline models (Alojail and Bhatia 2020). The most popular ensemble techniques for enhancing machine learning performance are bagging, boosting, and stacking. Table 1 describes the advantages and disadvantages of each.

There are several domains using ensemble learning methods to generalize machine learning techniques, such as natural language processing (NLP), internet of things (IoT), recommender systems, face recognition, information security, information retrieval, image retrieval, and intrusion detection system (Mohammed and Kora 2021; Forouzandeh et al. 2021; Yaman et al. 2018; Pashaei Barbin et al. 2020). Also, in sentiment analysis, many research studies have shown the superiority of the different ensemble learning methods over traditional machine learning classifiers. For example, the research efforts of Kanakaraj and Guddeti (2015); Prusa et al. (2015); Wang et al. (2014); Alrehili and Albalawi (2019); Sharma et al. (2018); Fersini et al. (2014); Perikos and Hatzilygeroudis (2016); Onan et al. (2016) applied a bagging method on a several of baseline classifiers such as (NB, SVM, KNN, LR, DT, ME) for English sentiment analysis. Also, the authors in Xia et al. (2011); Tsutsumi et al. (2007); Rodriguez-Penagos et al. (2013); Clark and Wicentwoski (2013); Li et al. (2010) applied two ensemble methods by voting and stacking based on NB, SVM and LR for

Table 1 Summary of ensemble methods

Ensemble methods	Advantage	Disadvantage
Bagging	- Ease of implementation and adapts. - Reducing Variance (Avoids Overfitting). - High performs on high-dimensional data. -Allowing weak learners to outperform strong learner -Robust against to noise or outliers data	-High Bias -Computationally Expensive -Loss of interpretability of the model
Boosting	-Reduces Variance. -Reduces Bias. -Handling of the missing data. - Ease of interpretation of the model	-Slower to train - Computationally Expensive -More Overfitting -The difficulty of scaling sequential training -Each classifier must correct the errors made by its predecessors
Stacking	-A deeper understanding of the data. -More Accurate -Less Variance -Less Bias -Used to ensemble a variety of strong learners	-More Overfitting - Time Complexity -The difficulty of interpreting the final model

English sentiment analysis. In addition, the authors in Da Silva et al. (2014); Xia et al. (2016); Fersini et al. (2016); Araque et al. (2017); Saleena (2018) applied majority voting based on several traditional classifiers such as SVM, RF, LR, NB, DT, and ME for English sentiment analysis. At the same time, several studies applied a stacking based on traditional classifiers for non-English sentiment analysis. For example, the authors in Lu and Tsou (2010); Li et al. (2012); Su et al. (2012) applied a stacking based on KNN, NB, SVM, and ME for Chinese reviews, the authors in Pasupulety et al. (2019) applied a stacking based on SVM and RF for India's reviews. In contrast, few studies applied ensemble learning techniques based on traditional classifiers of the Arabic language and its different dialects. For example, the authors in Saeed et al. (2022) applied a stacking based on SVM, NB, LR, DT, and KNN for Arabic sentiment analysis. But the authors in Oussous et al. (2018) applied a stacking based on SVM and ME for Moroccan tweets. On the other hand, ensemble-based deep learning models are a powerful alternative to traditional ensemble learning methods. Ensemble deep learning has shown excellent performance in sentiment analysis. For example, the researchers in Deriu et al. (2016); Akhtyamova et al. (2017) applied two ensemble methods by voting and stacking based on CNN for English sentiment analysis. Similarly, the work in Xu et al. (2016); Araque et al. (2017); Mohammadi and Shaverizade (2021); Haralabopoulos et al. (2020) applied voting and stacking based on LSTM and CNN for English sentiment analysis. However, the researchers in Heikal et al. (2018) applied voting based on CNN, GRU, and LSTM for Arabic sentiment analysis.

3 Proposed meta-ensemble deep learning approach

The meta-ensemble deep learning approach architecture consists of three layers, which are level-1, level-2, and level-3, as in Fig. 1. Level 1 represents the input layer, where each board of (M) models is trained independently using a unique training dataset and different deep architectures. Level 2 represents the meta-learner's hidden layer, in which each board model's prediction outputs in the previous layer are combined using a meta-learner. Level 3 represents the output meta-learner layer. At this level, the outputs of all predictions of the level-2 meta-learner are combined using the final level of the meta-learner to produce the final results. The proposed approach in abstract form can be seen as a general meta-neural network in which the first level is considered the input layer, level 2 is the hidden layer that acts as an activation function, and level 3 is the output layer.

3.1 Description of the proposed Algorithm

The formal semantics of the proposed training procedure of the proposed approach is shown in algorithm 1. The algorithm starts by randomly generating N equally-size samples from a training dataset $Data^{(0)}$. Each data sample $Data_i^{(0)}$=(train $_i^{(0)}$,test $_i^{(0)}$) is splitted into two parts; training and testing data. At the Baseline Learning procedure, the $Level-1$ learning models are generated by applying M BL_j Baseline Deep learners on each training dataset (train $_i^{(0)}$). As a result, we have n boards $C_i, 1 \leq i \leq n$ each containing M diverse baseline models $C_i = Model_{i1}, Model_{i2}, \dots, Model_{iM}$. For each test, $Test_i^{(0)} = (X^{(0)}, Y^{(0)})$, of the n data samples are used to create metadata $Data_i^{(1)}$ of the next level by stacking all the predicted output of each model $Model_i$. Each $Data_i^{(1)}$ in level-2 has $M + 1$ features: M features result from the prediction of the model in the board C_i on the $test^{(}0)$, and one extra feature represents the class label $Y^{(}0)$. In $Level-2$ once metadata has been generated, a set $ShallowClf$ of n shallow meta classifier is used to generate the models of Level-2. Following the creation of Level-2 models, test $_i^{(1)} = (X^{(a)}, Y^{(1)})$ are utilized to construct top the final meta data of $Level-3$. Likewise the previous level, the top metadata are generated in two steps. The first step generates $Data_i^{(1)}$ of $n + 1$ features results from the predictions of Level-2 models on $X^{(1)}$ and target class $Y^{(1)}$. In the next step, we construct $Data_i^{(1)}$ to form the final metadata. A Final meta learner is utilized to learn those top metadata in the Level-3 learning phase.

Procedure 1 Proposed Multi-level Training Algorithm

```
procedure GENERATING Input: Data
    $Generate\ Data_i^{(0)} = Train_i^{(0)} \bigcup Test_i^{(0)} = (X_i^{(0)}, Y_i^{(0)})), 1 \leq i \leq n$
procedure BASELINE LEARNING: $Level1$
    $BaseModels = \{BL_1, BL_2, \ldots, BL_M\} a\ group\ of\ M\ Baseline\ Deep\ learners$
    for $each\ Train_i^{(0)}, 1 \leq i \leq n$ do
        for $each\ BL_j \in BaseModels, 1 \leq j \leq M$ do
            $Model_{ij} \leftarrow fit\ (BL_j, Train_i^{(0)}), 1 \leq j \leq M$
        $C_i \leftarrow \{Model_{i1}, Model_{i2}, \ldots, Model_{iM}\}, 1 \leq i \leq n$
    for $each\ Model_{ij} \in C_i, 1 \leq i \leq n, 1 \leq j \leq M$ do
        $y_{ij}^{(1)} \leftarrow Model_{ij}(X_i^{(0)}), 1 \leq j \leq M$
        $Data_i^{(1)} \leftarrow FeatureStacking([y_{i1}^{(1)}, y_{i2,}^{(1)} \ldots, y_{ik}^{(1)}, Y_i^{(0)}]), 1 \leq i \leq n$
procedure LEARNING: LEVEL2
    $Divide\ Data_i^{(1)} = Train_i^{(1)} \bigcup (Test_i^{(1)} = (X_i^{(1)}, Y_i^{(1)})), 1 \leq i \leq n$
    $ShallowClf = \{sh_1, sh_2, \ldots, sh_n\} a\ group\ of\ n\ shallow\ learners$
    for $each\ Train_i^{(1)}, 1 \leq i \leq n$ do
        $shModel_j \leftarrow fit\ (sh_i, Train_i^{(1)}), 1 \leq j \leq n$
    $shallowModels \leftarrow \{shModel_1, shModel_2, \ldots, shModel_n\}$
procedure FINAL LEVEL LEARNING: LEVEL3
    for $each\ Test_i^{(1)} = (X_i^{(1)}, Y_i^{(1)}), 1 \leq i \leq n$ do
        for $each\ shModel_j \in shallowModels$ do
            $y_{ij}^{(2)} \leftarrow predict\ shModel_j(X_i^{(1)})$
        $Data_i^{(2)} \leftarrow PredictedStacking[y_{i1}^{(2)}, y_{i2}^{(2)}, \ldots, y_{in}^{(2)}, Y_i^{(1)}], 1 \leq i \leq n$
    $TopMetaData = stacked([Data_1^{(2)}, Data_2^{(2)}, \ldots, Data_n^{(2)}]^T)$
    $FinalClassifier \leftarrow is\ top\ level\ classifier$
    $FinalModel \leftarrow fit(FinalClassifier, TopMetaData)$
```

4 Experiment results

This section describes the benchmark datasets used for sentiment analysis, the selection of baseline deep models, and shallow meta-classifiers in the framework of the proposed meta-ensemble deep learning approach scheme.

4.1 Description of benchmark datasets

To evaluate the extended meta-ensemble deep learning approach, we selected six sentiment benchmark datasets for conducting the experiments based on English, Arabic, and different dialects: We propose the first dataset called “Arabic-Egyptian corpus 2”, which made up of 40,000 annotated tweets from the corpus (Mohammed and Kora 2019), and another extension of 10 K tweets which is available in Kora and Mohammed (2022). The later extension consists of 5k positive and 5k negative tweets from the Arabic language and the Egyptian dialect. The second dataset includes tweets in the Saudi dialect related to distance learning during the Covid19 pandemic (Aljabri et al. 2021). It contains a total of 1675 tweets, which includes more positive tweets than negative tweets. The third dataset is ASTD (Nabil et al. 2015). It contains about 10K Arabic tweets from different dialects and is classified into 797 positive and 1682 negative (Table 2). Tweets were annotated as positive, neutral, negative, and mixed. The fourth dataset is ArSenTD-LEV (Al-Laith and Shahbaz 2021). It contains 4000 tweets from countries in the Levant Region, such as Jordan, Palestine, Lebanon and Syria. The fifth dataset is Movie Reviews (Koh et al. 2010). It contains 10,662 reviews, divided into 5331 negative and 5331 positives. The sixth dataset is the Twitter US Airline Sentiment dataset (Rane and Kumar 2018). Table 3 summarizes the characteristics of different benchmark datasets for sentiment analysis. It contains 14,600 customer tweets from six airlines in the US, including negative, positive, and neutral sentiments. In general, the textual data was preprocessed using one-hot encoding or word-embedding (Lai et al. 2016), as an initial layer before training the network. Only the positive and negative binary sentiment polarity labels are used for each dataset, and the other polarity labels are neglected. In our experiments, we divided each benchmark dataset into training and validation test sets with a ratio of (80%, 20%). In addition, we divided each benchmark dataset into eight partitions.

Table 2 Applications of ensemble learning to sentiment classification

Approach	Papers	Baseline classifiers	Ensemble method	Languages	Dataset
TEL	Wilson et al. (2006)	DT	Boosting	English	MPQA Corpus (Wiebe et al. 2005)
	Tsutsumi et al. (2007)	SVM, ME	Stacking	English	Movie Review (Chaovalit and Zhou 2005)
	Li et al. (2010)	SVM, LR	Voting	English	Amazon.com. (Rushdi-Saleh et al. 2011)
	Lu and Tsou (2010)	NB, ME, SVM	Stacking	Chinese	Reviews (Seki et al. 2008)
	Xia et al. (2011)	NB, ME, SVM	Stacking	English	Movie Review (Chen et al. 2012)
	Li et al. (2012)	SVM, KNN	Stacking	Chinese	Reviews (Seki et al. 2008)
	Su et al. (2012)	ME, SVM	Voting, Stacking	Chinese	Reviews (Seki et al. 2008)
	Rodriguez-Penagos et al. (2013)	SVM	Voting	English	SemEval (Dzikovska et al. 2013)
	Clark and Wicentwoski (2013)	NB	Voting	English	SemEval (Nakov et al. 2016)
	Fersini et al. (2014)	ME, SVM, NB	Voting,Bagging	English	Product Reviews Pang and Lee (2005)
	Da Silva et al. (2014)	SVM, RF, LR	Voting	English	Tweets Saif et al. (2013)
	Wang et al. (2014)	SVM, KNN, DT, ME, NB	Bagging,Boosting	English	Movie Reviews (Chaovalit and Zhou 2005)
	Kanakaraj and Guddeti (2015)	NB, SVM	Bagging,Boosting	English	Movie Review (Chen et al. 2012)
	Prusa et al. (2015)	KNN, SVM, LR	Bagging,Boosting	English	sentiment140 Corpus (Go et al. 2009)
	Xia et al. (2016)	SVM, LR	Voting	English	Amazon.com. (Rushdi-Saleh et al. 2011)
	Onan et al. (2016)	BLR, NB, LDA,LR, SVM	Stacking,AdaBoost, Bagging	English	Tweets (Whitehead and Yaeger 2009)
	Fersini et al. (2016)	NB, DT, SVM	Voting	English	Movie Reviews (Chen et al. 2012)
	Perikos and Hatzilygeroudis (2016)	NB, ME	Bagging	English	Posts (Cambria et al. 2013)
	Araque et al. (2017)	NB, ME, SVM	Voting	English	Movie Reviews (Chen et al. 2012)
	Oussous et al. (2018)	MNB, SVM, ME	Voting, Stacking	Moroccan	Tweets (Tratz et al. 2013)
	Saleena (2018)	SVM, RF, NB, LR	Voting	English	Sentiment140 Corpus (Go et al. 2009), Tweets (Speriosu et al. 2011)
	Sharma et al. (2018)	SVM	Bagging	English	Movie Reviews (Chen et al. 2012)
	Pasupulety et al. (2019)	SVM, RF	Stacking	Indian	NSE (Kumar and Misra 2018)
	Saeed et al. (2022)	SVM, NB, LR, DT, KNN	Voting, Stacking	Arabic	Corpus (Li et al. 2011)

Table 2 (continued)

Approach	Papers	Baseline classifiers	Ensemble method	Languages	Dataset
EDL	Deriu et al. (2016)	CNN	Stacking	English	SemEval (Bethard et al. 2016)
	Xu et al. (2016)	CNN, LSTM	Voting	English	SemEval (Dzikovska et al. 2013)
	Akhtyamova et al. (2017)	CNNs	Voting	English	Reviews (Karimi et al. 2015)
	(Araque et al. 2017)	CNN, LSTM, GRU	Voting, Stacking	English	Movie reviews (Chen et al. 2012)
	(Heikal et al. 2018)	CNN, LSTM	Voting	Arabic	ASTD (Nabil et al. 2015)
	Haralabopoulos et al. (2020)	LSTM, GRU, CNN, RCNN, DNN	Voting, Stacking	English	Comments (van Aken et al. 2018), SemEval (Bethard et al. 2016)
	(Mohammadi and Shaverizade 2021)	CNN, LSTM, GRU, Bi_LSTM	Stacking	English	SemEval (Bethard et al. 2016)

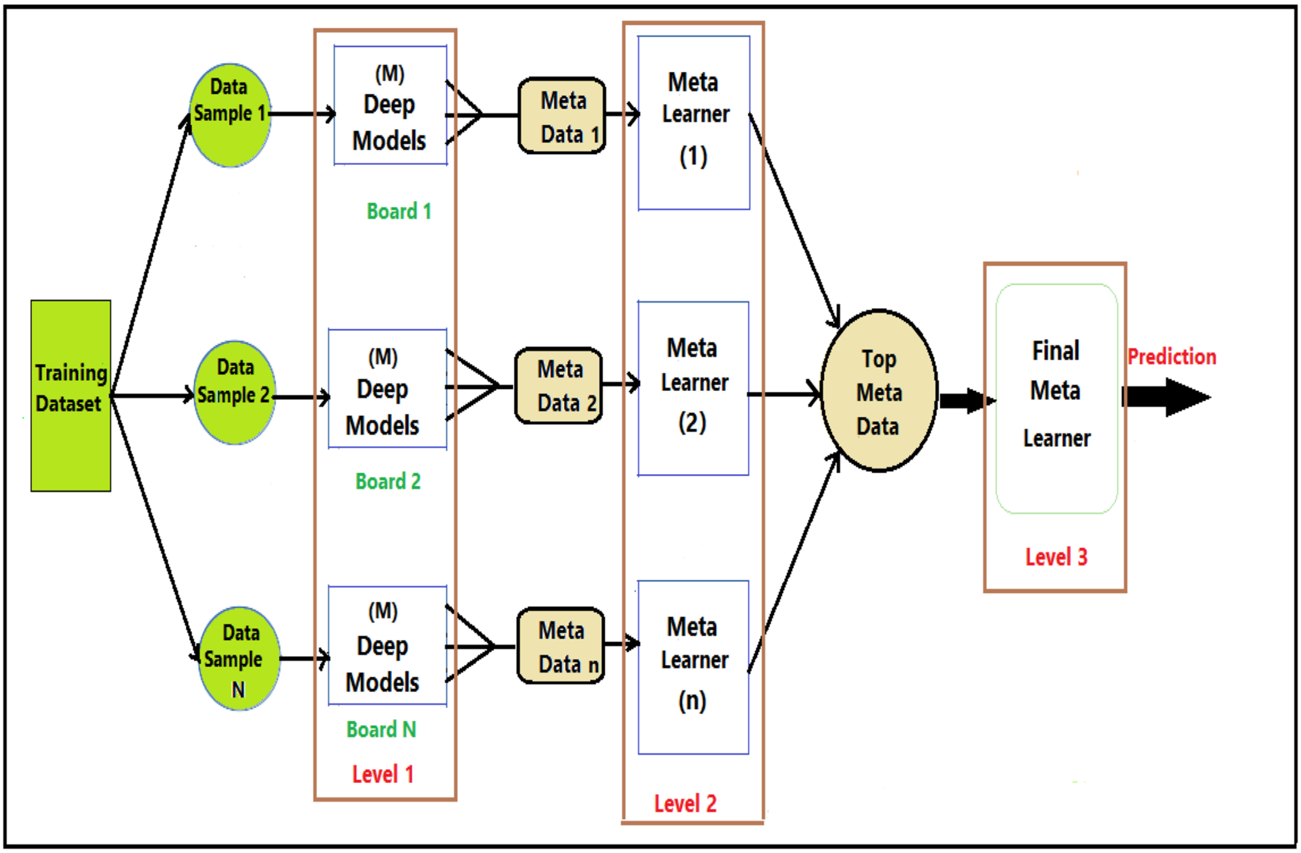

Fig. 1 The general architecture of the proposed meta-ensemble deep learning approach

4.2 Baseline deep learning models

To enhance the performance of predictions in sentiment analysis through the proposed meta-ensemble deep learning approach, we first need to build a set of deep learning models that form the baseline classifiers of the proposed meta-ensemble deep learning approach for each benchmark dataset. Three deep baseline models are proposed in this research: Long Short-Term Memory (LSTM) is the first

Table 3 Distribution of the different benchmark dataset

Dataset	Data types	Sentiment classes	Positive count	Negative count	Total count
1-Arabic-Egyptian Corpus (Mohammed and Kora 2019; Kora and Mohammed 2022)	Egyptian dialects, MSA	2	25k	25k	50k
2-Saudi Arabia Tweets (Aljabri et al. 2021)	Dialects Tweets	2	1002	673	1675
3-ASTD (Nabil et al. 2015)	Dialects Tweets	4	797	1682	10,006
4-ArSenTD-LEV (Al-Laith and Shahbaz 2021)	Dialects Tweets	5	835	1253	4,000
5-Movie Reviews (Koh et al. 2010)	English Reviews	2	5331	5331	10,662
6-Twitter US Airline Sentiment (Rane and Kumar 2018)	English Tweets	3	2310	8797	14,601

Table 4 Configurations of baseline deep learning models

Models	Configuration value
GRU	GRU layer= 1 or 2
	GRU size= 256
LSTM	LSTM layer= 1 or 2
	LSTM size= 256
CNN	No. of filters= 32
	Filters size= 16
	Vocab size= 10,000

baseline deep model utilized in our evaluation (Mohammed and Kora 2019). The LSTM model is a well-known architecture for representing sequential data. It was designed better to capture long-term dependencies than the recurrent neural network model. Three gates comprise LSTM architecture: the input gate, the forget gate, and the output gate. The Gated recurrent unit (GRU) is the next baseline deep model (Pan et al. 2020). The GRU model is comparable to the LSTM model, except it contains fewer parameters. GRU comprises of two gates: the reset gate and the update gate. The Convolutional Neural Network Model (CNN) is the third baseline deep model (Abdulnabi et al. 2015). The CNN model is a feedforward neural network consisting of one or more convolutional layers and a fully connected layer, which also includes a pooling layer for integration. In general, each deep baseline model is trained on different hyperparameters. Table 4 shows the configurations of baseline deep learning models. Table 5 shows the accuracy of each data split within each dataset and the average accuracy of each baseline deep model in each dataset. It should be mentioned that the experimental results reveal that the highest average accuracy obtained in the first dataset of Arabic-Egyptian Corpus is 89.38% of the LSTM model. Also, the highest average accuracy obtained in the second dataset of Saudi Arabia Tweets is 65.38% of the LSTM2 model. In addition, the highest average accuracy obtained in the third ASTD dataset is 71.6% of the LSTM model. Moreover, the highest average accuracy obtained in the fourth ArSenTD-LEV dataset is 76.2% of the LSTM model. Additionally, the highest average accuracy obtained in the fifth dataset of the Movie Reviews dataset is 78.03% of the LSTM1 model. Finally, the highest average accuracy obtained in the Twitter US Airline Sentiment dataset's sixth dataset is 80.05% of the LSTM1 model. In the conducted experiments, 114 deep baseline models in all have been trained. In addition, the sizes of the baseline models vary on each dataset. In Saudi Arabia, tweets, Movie Reviews, and Twitter US Airline Sentiment are 4 deep baseline models, while ASTD and ArSenTD-LEV are 3 deep baseline models.

4.3 Meta-ensemble classifiers

To combine the trained baseline deep models within the boards of models, we use a set of shallow meta-classifiers that include Support Vector Machines (SVM), Gradient Boosting (GB), Naive Bayes (NB), Random Forest (RF), Logistic Regression (LG) as top surface meta learners. Table 6 describes the accuracy results of the proposed clustering method in each dataset. In the first dataset of Arabic-Egyptian Corpus, the results indicate that the ensemble with SVM classifier achieved the best accuracy in both hard and soft prediction with a score of 92.6% and 93.2%, respectively. In the second dataset of Saudi Arabian tweets, the results indicate that the ensemble with the SVM classifier achieved the best accuracy in the hard prediction of 69.9%. In contrast, the ensemble with both the SVM and LG classifier achieved the best soft prediction accuracy with a score of 72.3%. In the third dataset of ASTD, the results indicate that both the ensemble with SVM and LG classifier achieved the best accuracy in hard prediction with a score of 75.9%. At the same time, the ensemble with the LG classifier achieved the best accuracy in soft prediction with a score of 77.6%. In the fourth dataset of ArSenTD-LEV, the results indicate that the ensemble with the SVM classifier achieved the best accuracy in hard prediction with a score of 80.4%. In contrast, the ensemble with the LG classifier achieved the best accuracy in soft prediction with a score of 83.2%. In

Table 5 Performance accuracy results of baseline deep classifiers in different datasets

		Split dataset								
Dataset	Baseline models	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	**AVG models** (%)
1-Mohammed and Kora (2019); Kora and Mohammed (2022)	GRU	**89.9**	**89.8**	89.2	**89.5**	**89.3**	88.8	88.2	89	89.21
	LSTM	89.7	**89.8**	**89.8**	89.1	89.2	**89**	**89.1**	**89.4**	**89.38**
	CNN	87.64	85.04	84.78	85.65	87.40	86	85	86.2	85.96
2-Aljabri et al. (2021)	GRU1	**63.1**	67.3	**66.2**	64.2	62.3	64.2	**65.4**	67.7	65.05
	LSTM1	61.9	61.9	60	65	64.6	68.5	65	**70.8**	64.71
	GRU2	60.8	**69.6**	60.4	61.5	63.1	66.9	64.6	61.5	63.55
	LSTM2	60.4	65.4	65	**66.2**	**66.5**	**70**	62.3	67.3	**65.38**
	CNN	–	–	–	–	–	–	–	–	–
3-Nabil et al. (2015)	GRU1	**73.1**	66.2	68.5	72.8	**74.1**	**72.3**	**72.8**	67.9	70.86
	LSTM1	72.1	**75.9**	**69.5**	**74.1**	71.5	69.2	69	**71.5**	**71.6**
	GRU2	–	–	–	–	–	–	–	–	–
	LSTM2	–	–	–	–	–	–	–	–	–
	CNN	68.2	70.4	67	68	68.9	71	70.6	68.2	69.03
4-Al-Laith and Shahbaz (2021)	GRU1	**74.5**	**76.4**	73.6	73.9	76.4	77.3	**78.5**	76.4	75.87
	LSTM1	73.3	75.8	**75.2**	**78.2**	75.2	**78.2**	75.8	**77.9**	**76.2**
	GRU2	–	–	–	–	–	–	–	–	–
	LSTM2	–	–	–	–	–	–	–	–	–
	CNN	70.5	76	65.5	75	71.3	77.3	75.5	66	72.13
5-Koh et al. (2010)	GRU1	68.9	**77.5**	76.6	75.4	71.2	75.3	74.8	**76.6**	74.53
	LSTM1	**82.6**	74.8	**79.4**	**81.9**	**81.7**	**76.4**	**82.7**	64.8	**78.03**
	GRU2	62.4	57.4	55.4	64.1	66.1	58.2	69	69.2	62.72
	LSTM2	71.9	67.9	62.4	68.4	54.8	66.8	65.9	74.9	66.62
	CNN	–	–	–	–	–	–	–	–	–
6-Rane and Kumar (2018)	GRU1	71.6	**78.6**	79.2	78.9	68.5	70.4	65.9	73.3	73.18
	LSTM1	**80.6**	78.4	**81.1**	**79.7**	**80.3**	**81.8**	**78.1**	**81.2**	**80.05**
	GRU2	70.6	66.4	70.8	68.2	63.3	67.7	64.4	63.9	66.82
	LSTM2	73.2	66.3	72.4	69.7	71.3	70.8	70.9	73.1	70.96
	CNN	–	–	–	–	–	–	–	–	–

The values in bold indicate superior results among the baseline models in each data split

Table 6 Performance Accuracy of the proposed Meta-Ensemble in different datasets

Dataset	Predictions	GB (%)	SVM (%)	NB (%)	LG (%)	RF (%)
1-Mohammed and Kora (2019); Kora and Mohammed (2022)	Hard	92	**92.6**	91.6	91.9	91.9
	Soft	91.8	**93.2**	92.2	92.3	90
2-Aljabri et al. (2021)	Hard	69.3	**69.9**	67.4	69.2	68.4
	Soft	71.2	**72.3**	69.8	**72.3**	71.8
3-Nabil et al. (2015)	Hard	74.1	**75.9**	72.3	**75.9**	74.1
	Soft	76.2	77.1	73.6	**77.6**	75.8
4-Al-Laith and Shahbaz (2021)	Hard	79.5	**80.4**	76.2	80.3	79.6
	Soft	81.4	82.3	79.1	**83.2**	81.4
5-Koh et al. (2010)	Hard	80.5	**80.9**	79.3	80.5	80.5
	Soft	82.4	**83.9**	80.5	83.8	82.1
6-Rane and Kumar (2018)	Hard	82.1	**82.9**	80.3	81.8	82.2
	Soft	**85.3**	85.1	81.9	85.1	84.9

The values in bold indicate superior results among the baseline models in each data split

Table 7 Summary of accuracy

Benchmarks	AVG Baseline models	High AVG Baseline models	Meta-Ensemble

1-Mohammed and Kora (2019); Kora and Mohammed (2022)	GRU= 89.52%
	LSTM= 89.54%
	CNN= 86.10%
		LSTM= 89.54%	SVM=**93.2%** (Soft)

2-Aljabri et al. (2021)	GRU1= 65.05%
	LSTM1= 64.71%
	GRU2= 63.55%
	LSTM2= 65.38%
		LSTM2= 65.38%	SVM=**72.3%** (Soft)

3-Nabil et al. (2015)	GRU= 70.86%
	LSTM= 71.6%
	CNN= 69.03%
		LSTM= 71.6%	LG=**77.6%** (Soft)

4-Al-Laith and Shahbaz (2021)	GRU= 75.87%
	LSTM= 76.2%
	CNN= 72.13%
		LSTM= 76.2%	LG=**83.2%** (Soft)

5-Koh et al. (2010)	GRU1= 74.53%
	LSTM1= 78.03%
	GRU2= 62.72%
	LSTM2= 66.62%
		LSTM1= 78.03%	SVM=**83.9%** (Soft)

6-Rane and Kumar (2018)	GRU1= 73.18%
	LSTM1=80.05%
	GRU2= 66.82%
	LSTM2=70.96%
		LSTM1=80.05%	GB=**85.3%** (Soft)

The values in bold indicate superior results among the meta classifiers in each data split

the fifth Movie Reviews dataset, the results indicate that the ensemble with the SVM classifier achieved the best accuracy in both hard and soft prediction with a score of 80.9% and 83.9%, respectively. In the sixth dataset of Twitter US Airline Sentiment, the results indicate that the ensemble with the SVM classifier achieved the best accuracy in hard prediction with a score of 82.9%. At the same time, the ensemble with the GB classifier achieved the best accuracy in soft prediction with a score of 85.3%. Table 7 compares the highest accuracy results of the average baseline deep models with the highest accuracy results of meta-ensemble classifiers in each dataset. It can be noted that the highest average accuracy was obtained in the proposed meta-ensemble in the different datasets in soft prediction. Also, it can be noted that the highest average accuracy obtained in baseline deep models in the different datasets is the LSTM model than in the other networks. In general, it can be noted that different meta-ensemble classifiers show better performance for the final prediction. It can also be noted that using 5-fold cross-validation on the predictions of deep baseline models, SVM is shown as the most frequent best combiner to fuse the boards of models in the level-1 with 93.2%, 72.3% and 83.9% in each of the Arabic-Egyptian Corpus, Saudi Arabia Tweets and Movie Reviews datasets, respectively. In addition, LG is shown as the most frequent best combiner to fuse the boards of models in level-1 with 77.6% and 83.2% in both the ASTD and ArSenTD-LEV datasets, respectively. Finally, GB is considered the most frequent best combiner to fuse the models' boards in the level-1 at 85.3% in the Twitter US Airline Sentiment datasets.

5 Conclusion

Deep learning models have shown great success in sentiment analysis in the literature. However, modeling an effective deep learning model requires great effort due to finding the best architecture of the neural network and the best configuration of hyperparameters. An approach for tackling these limitations is using the ensemble methods. The key idea of the ensemble is to produce a powerful learner using a combination of weak learners. Thus, in this research paper, we proposed a meta-ensemble deep learning approach to improve the performance of sentiment analysis. This proposed approach combines the predictions of several groups of deep models using three levels of the meta-learning method. Also, we proposed the benchmark dataset "Arabic-Egyptian

Corpus 2”. This corpus comprises 10,000 annotated tweets written in colloquial Arabic on various topics. This corpus is added to the original version in Mohammed and Kora (2019) that contains 40K annotated tweets. We conducted several experiments on six public benchmark datasets for sentiment analysis involving several languages and dialects to test and evaluate the performance of the proposed meta-ensemble deep learning approach. We trained sets of baseline classifiers (GRU, LSTM, and CNN) on each benchmark dataset, and their best model was compared with the proposed meta-ensemble deep learning approach. In particular, we have trained 114 deep models and performed a comparison on five different shallow meta-classifiers to ensemble those models. The experimental results revealed that the meta-ensemble deep learning approach effectively outperforms all six benchmark datasets’ baseline deep learning models. Also, the experiments suggested that the meta-learners work better when the predictions of the involved layers are of the form probability distribution. In summary, the proposed ensemble approach uses parallel ensemble techniques where baseline learners are generated simultaneously, as there is no data dependency and the fusion methods depend on the meta-learning method. However, our proposed approach has some challenges and limitations, such as determining the appropriate number of baseline models and selecting baseline models that can be relied upon to generate the best predictions from each dataset when designing our meta-ensemble deep learning approach from scratch. Also, the difficulty of computing time complexity is added when the amount of available data grows exponentially. In addition, the issue of multi-label classification raises many problems, such as overfitting and the curse of dimensionality, in the case of high dimensionality of data. Handling a multi-class problems worth investigating in case of multi-level ensemble. Also, transformer models recently received more attention in NLP tasks. It is worth investigating the impact of ensemble learning with transformers with full extensive experiments.

Author contributions Paper is written by AM and RK Paper is reviewed by AM.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abdulnabi AH, Wang G, Lu J, Jia K (2015) Multi-task cnn model for attribute prediction. IEEE Trans Multimedia 17(11):1949–1959

Ahmed S, Pasquier M, Qadah G (2013) Key issues in conducting sentiment analysis on arabic social media text. In: 2013 9th International conference on innovations in information technology (IIT), pp 72–77. IEEE

van Aken B, Risch J, Krestel R, Löser (2018) A challenges for toxic comment classification: an in-depth error analysis. In: ALW

Akhtyamova L, Ignatov A, Cardiff J (2017) A large-scale cnn ensemble for medication safety analysis. In: International conference on applications of natural language to information systems, pp 247–253. Springer

Al-Laith A, Shahbaz M (2021) Tracking sentiment towards news entities from arabic news on social media. Future Gener Comput Syst 118:467–484

Aljabri M, Chrouf SMB, Alzahrani NA, Alghamdi L, Alfehaid R, Alqarawi R, Alhuthayfi J, Alduhailan N (2021) Sentiment analysis of arabic tweets regarding distance learning in saudi arabia during the covid-19 pandemic. Sensors 21(16):5431

Alojail M, Bhatia S (2020) A novel technique for behavioral analytics using ensemble learning algorithms in e-commerce. IEEE Access 8:150072–150080

Alomari KM, ElSherif HM, Shaalan K (2017) Arabic tweets sentimental analysis using machine learning. In: International conference on industrial, engineering and other applications of applied intelligent systems, pp 602–610. Springer

Alrehili A, Albalawi K (2019) Sentiment analysis of customer reviews using ensemble method, pp 1–6

Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA (2017) Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Syst Appl 77:236–246

Baly R, El-Khoury G, Moukalled R, Aoun R, Hajj H, Shaban KB, El-Hajj W (2017) Comparative evaluation of sentiment analysis methods across arabic dialects. Procedia Comput Sci 117:266–273

Bethard S, Savova G, Chen WT, Derczynski L, Pustejovsky J, Verhagen M (2016) Semeval-2016 task 12: clinical tempeval. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp 1052–1062

Cambria E, Das D, Bandyopadhyay S, Feraco A, et al (2017) A practical guide to sentiment analysis

Cambria E, Schuller B, Xia Y, Havasi C (2013) New avenues in opinion mining and sentiment analysis. IEEE Intell Syst 28(2):15–21

Chan S, Reddy V, Myers B, Thibodeaux Q, Brownstone N, Liao W (2020) Machine learning in dermatology: current applications, opportunities, and limitations. Dermatol Therapy 10(3):365–386

Chaovalit P, Zhou L (2005) Movie review mining: a comparison between supervised and unsupervised classification approaches. In: Proceedings of the 38th annual Hawaii international conference on system sciences, pp 112c–112c. IEEE

Chen L, Wang W, Nagarajan M, Wang S, Sheth A (2012) Extracting diverse sentiment expressions with target-dependent polarity from twitter. In: Proceedings of the international AAAI conference on web and social media, vol 6, pp 50–57

Chen Y, Yuan J, You Q, Luo J (2018) Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In: 2018 ACM Multimedia conference on multimedia conference, pp 117–125. ACM

Cho SB, Won HH (2003) Machine learning in dna microarray analysis for cancer classification. In: Proceedings of the First Asia-Pacific bioinformatics conference on bioinformatics 2003-volume 19, pp 189–198

Clark S, Wicentwoski R (2013) Swatcs: combining simple classifiers with estimated accuracy. In: Second joint conference on lexical and computational semantics (* SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), pp 425–429

Collobert R, Weston J (2008) A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th international conference on machine learning, pp 160–167

Da Silva NF, Hruschka ER, Hruschka ER Jr (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66:170–179

Deriu J, Gonzenbach M, Uzdilli F, Lucchi A, Luca VD, Jaggi M (2016) Swisscheese at semeval-2016 task 4: sentiment classification using an ensemble of convolutional neural networks with distant supervision. In: Proceedings of the 10th international workshop on semantic evaluation, CONF, pp 1124–1128

Duwairi RM, Marji R, Sha'ban N, Rushaidat S (2014) Sentiment analysis in arabic tweets. In: 2014 5th International conference on information and communication systems (ICICS), pp 1–6. IEEE

Dzikovska MO, Nielsen RD, Brew C, Leacock C, Giampiccolo D, Bentivogli L, Clark P, Dagan I, Dang HT (2013) Semeval-2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. North Texas State Univ Denton, Tech. rep

Fersini E, Messina E, Pozzi FA (2014) Sentiment analysis: Bayesian ensemble learning. Decis Support Syst 68:26–38

Fersini E, Messina E, Pozzi FA (2016) Expressive signals in social media languages to improve polarity detection. Inf Process Manag 52(1):20–35

Forouzandeh S, Berahmand K, Rostami M (2021) Presentation of a recommender system with ensemble learning and graph embedding: a case on movielens. Multimedia Tools Appl 80(5):7805–7832

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N project report, Stanford 1(12), 2009

Graves A (2012) Long short-term memory. Supervised sequence labelling with recurrent neural networks, pp 37–45

Habimana O, Li Y, Li R, Gu X, Yu G (2020) Sentiment analysis using deep learning approaches: an overview. Sci China Inf Sci 63(1):1–36

Haralabopoulos G, Anagnostopoulos I, McAuley D (2020) Ensemble deep learning for multilabel binary classification of user-generated content. Algorithms 13(4):83

Heikal M, Torki M, El-Makky N (2018) Sentiment analysis of arabic tweets using deep learning. Procedia Comput Sci 142:114–122

Kanakaraj M, Guddeti RMR (2015) Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In: Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015), pp 169–170. IEEE

Karimi S, Metke-Jimenez A, Kemp M, Wang C (2015) Cadec: a corpus of adverse drug event annotations. J Biomed Inform 55:73–81

Koh NS, Hu N, Clemons EK (2010) Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. Electron Commer Res Appl 9(5):374–385

Kora R, Mohammed A (2022) Arabic-Egyptian Corpus 2. https://doi.org/10.7910/DVN/UPGJCV

Kulkarni NH, Srinivasan G, Sagar B, Cauvery N (2018) Improving crop productivity through a crop recommendation system using ensembling technique. In: 2018 3rd International conference on computational systems and information technology for sustainable solutions (CSITSS), pp 114–119. IEEE

Kumar G, Misra AK (2018) Commonality in liquidity: evidence from India's national stock exchange. J Asian Econ 59:1–15

Kumar V, Aydav PSS, Minz S (2021) Multi-view ensemble learning using multi-objective particle swarm optimization for high dimensional data classification. J King Saud Univ-Comput Inf Sci

Lai S, Liu K, He S, Zhao J (2016) How to generate a good word embedding. IEEE Intell Syst 31(6):5–14

Le NQK, Yapp EKY, Yeh HY (2019) Et-gru: using multi-layer gated recurrent units to identify electron transport proteins. BMC Bioinform 20(1):1–12

Li FH, Huang M, Yang Y, Zhu X (2011) Learning to identify review spam. In: Twenty-second international joint conference on artificial intelligence

Li S, Lee SY, Chen Y, Huang CR, Zhou G (2010) Sentiment classification and polarity shifting. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), pp 635–643

Li W, Wang W, Chen Y (2012) Heterogeneous ensemble learning for Chinese sentiment classification. J Inf Comput Sci 9(15):4551–4558

Lu B, Tsou BK (2010) Combining a large sentiment lexicon and machine learning for subjectivity classification. In: 2010 international conference on machine learning and cybernetics, vol 6, pp 3311–3316. IEEE

Mejova Y (2009) Sentiment analysis: an overview. University of Iowa, Computer Science Department

Mohammadi A, Shaverizade A (2021) Ensemble deep learning for aspect-based sentiment analysis. Int J Nonlinear Anal Appl 12(Special Issue):29–38

Mohammed A, Kora R (2019) Deep learning approaches for arabic sentiment analysis. Soc Netw Anal Min 9(1):52

Mohammed A, Kora R (2021) An effective ensemble deep learning framework for text classification. J King Saud Univ-Comput Inf Sci

Moitra D, Mandal RK (2019) Automated ajcc staging of non-small cell lung cancer (nsclc) using deep convolutional neural network (cnn) and recurrent neural network (rnn). Health Inf Sci Syst 7(1):1–12

Nabil M, Aly M, Atiya A (2015) Astd: Arabic sentiment tweets dataset. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 2515–2519

Nakov P, Rosenthal S, Kiritchenko S, Mohammad SM, Kozareva Z, Ritter A, Stoyanov V, Zhu X (2016) Developing a successful semeval task in sentiment analysis of twitter and other social media texts. Lang Resour Eval 50(1):35–65

Naresh A, Venkata Krishna P (2021) An efficient approach for sentiment analysis using machine learning algorithm. Evol Intel 14(2):725–731

Onan A, Korukoğlu S, Bulut H (2016) A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst Appl 62:1–16

Oussous A, Lahcen AA, Belfkih S (2018) Improving sentiment analysis of moroccan tweets using ensemble learning. In: International conference on big data, cloud and applications, pp 91–104. Springer

Pan M, Zhou H, Cao J, Liu Y, Hao J, Li S, Chen CH (2020) Water level prediction model based on gru and cnn. IEEE Access 8:60090–60100

Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: ACL

Pashaei Barbin J, Yousefi S, Masoumi B (2020) Efficient service recommendation using ensemble learning in the internet of things (iot). J Ambient Intell Humaniz Comput 11(3):1339–1350

Pasupulety U, Anees AA, Anmol S, Mohan BR (2019) Predicting stock prices using ensemble learning and sentiment analysis. In: 2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE), pp 215–222. IEEE

Perikos I, Hatzilygeroudis I (2016) Recognizing emotions in text using ensemble of classifiers. Eng Appl Artif Intell 51:191–201

Pontiki M, Galanis D, Papageorgiou H, Androutsopoulos I, Manandhar S, Mohammad AS, Al-Ayyoub M, Zhao Y, Qin B, De Clercq O, et al (2016) Semeval-2016 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp 19–30

Prusa J, Khoshgoftaar TM, Dittman DJ (2015) Using ensemble learners to improve classifier performance on tweet sentiment data. In: 2015 IEEE international conference on information reuse and integration, pp 252–257. IEEE

Rane A, Kumar A (2018) Sentiment classification system of twitter data for us airline service analysis. In: 2018 IEEE 42nd annual computer software and applications conference (COMPSAC), vol 1, pp 769–773. IEEE

Rodriguez-Penagos C, Atserias J, Codina-Filba J, García-Narbona D, Grivolla J, Lambert P, Saurí R (2013) Fbm: combining lexicon-based ml and heuristics for social media polarities. In: Second joint conference on lexical and computational semantics (*SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013), pp 483–489

Rojas-Barahona LM (2016) Deep learning for sentiment analysis. Lang Linguist Compass 10(12):701–719

Rushdi-Saleh M, Martín-Valdivia MT, Ureña-López LA, Perea-Ortega JM (2011) Oca: opinion corpus for arabic. J Am Soc Inform Sci Technol 62(10):2045–2054

Saeed RM, Rady S, Gharib TF (2022) An ensemble approach for spam detection in arabic opinion texts. J King Saud Univ-Comput Inf Sci 34(1):1407–1416

Sagi O, Rokach L (2018) Ensemble learning: a survey. Wiley Interdiscip Rev: Data Min Knowl Discovery 8(4):e1249

Saif H, Fernandez M, He Y, Alani H (2013) Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold

Saleena N et al (2018) An ensemble classification system for twitter sentiment analysis. Procedia Comput Sci 132:937–946

Seki Y, Evans DK, Ku LW, 0001, L.S., Chen HH, Kando N (2008) Overview of multilingual opinion analysis task at ntcir-7. In: NTCIR, pp 185–203. Citeseer

Shahzad RK, Lavesson N (2013) Comparative analysis of voting schemes for ensemble-based malware detection. J Wirel Mobile Netw Ubiquitous Comput Depend Appl 4(1):98–117

Sharma S, Srivastava S, Kumar A, Dangi A (2018) Multi-class sentiment analysis comparison using support vector machine (svm) and bagging technique-an ensemble method. In: 2018 International conference on smart computing and electronic enterprise (ICSCEE), pp 1–6. IEEE

Shipp CA, Kuncheva LI (2002) Relationships between combination methods and measures of diversity in combining classifiers. Inf Fusion 3(2):135–148

Shoukry A, Rafea A (2012) Sentence-level arabic sentiment analysis. In: 2012 International conference on collaboration technologies and systems (CTS), pp 546–550. IEEE

Speriosu M, Sudan N, Upadhyay S, Baldridge J (2011) Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the first workshop on unsupervised learning in NLP, pp 53–63

Stamatatos E, Widmer G (2002) Music performer recognition using an ensemble of simple classifiers. In: ECAI, pp 335–339

Su Y, Zhang Y, Ji D, Wang Y, Wu H (2012) Ensemble learning for sentiment classification. In: Workshop on Chinese lexical semantics, pp 84–93. Springer

Tan KL, Lee CP, Lim KM, Anbananthen KSM (2022) Sentiment analysis with ensemble hybrid deep learning model. IEEE Access 10:103694–103704

Tasci E, Uluturk C, Ugur A (2021) A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. Neural Comput Appl, pp 1–15

Tratz S, Briesch D, Laoudi J, Voss C (2013) Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija. In: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, pp 135–139

Tsutsumi K, Shimada K, Endo T (2007) Movie review classification based on a multiple classifier. In: Proceedings of the 21st pacific Asia conference on language, information and computation, pp 481–488

Tuysuzoglu G, Birant D, Pala A (2018) Ensemble methods in environmental data mining. Sch Environ Sci, pp 1–16

Wagh R, Punde P (2018) Survey on sentiment analysis using twitter dataset. In: 2018 Second international conference on electronics, communication and aerospace technology (ICECA), pp 208–211. IEEE

Wang G, Sun J, Ma J, Xu K, Gu J (2014) Sentiment classification: the contribution of ensemble learning. Decis Support Syst 57:77–93

Wang XY, Zhang BB, Yang HY (2013) Active svm-based relevance feedback using multiple classifiers ensemble and features reweighting. Eng Appl Artif Intell 26(1):368–381

Whitehead M, Yaeger L (2009) Building a general purpose cross-domain sentiment mining model. In: 2009 WRI world congress on computer science and information engineering, vol 4, pp 472–476. IEEE

Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 39(2):165–210

Wilson T, Wiebe J, Hwa R (2006) Recognizing strong and weak opinion clauses. Comput Intell 22(2):73–99

Xia R, Xu F, Yu J, Qi Y, Cambria E (2016) Polarity shift detection, elimination and ensemble: a three-stage model for document-level sentiment analysis. Inf Process Manag 52(1):36–45

Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci 181(6):1138–1152

Xu S, Liang H, Baldwin T (2016) Unimelb at semeval-2016 tasks 4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016), pp 183–189

Yadav A, Vishwakarma DK (2020) Sentiment analysis using deep learning architectures: a review. Artif Intell Rev 53(6):4335–4385

Yaman MA, Subasi A, Rattay F (2018) Comparison of random subspace and voting ensemble machine learning methods for face recognition. Symmetry 10(11):651

Zhang C, Ma Y (2012) Ensemble machine learning: methods and applications. Springer