Bukti korespondensi antara Penulis 1 dengan Correspondent Author.

1. Original Manuscript (5 October 2021)
2. First review: Major revision (7 March 2022)
3. Reminder from Transactions on Internet Technology Editorial Office (4 April 2022)
4. First Response to the reviewers' and First Revised manuscript files are sent to the Correspondent Author (12 April 2022), and submitted to the Transactions on Internet Technology journal (17 Apr 2022)
    a. First response to the reviewers'
    b. First Revised manuscript with brown font for the revision
5. Second review: Minor revision (22 June 2022)
6. Second Response to the reviewers' and Second Revision manuscript files are sent to the Correspondent Author (30 June 2022)
    a. Second response to the reviewers'
    b. Second Revision manuscript with brown font for the revision
7. Paper accepted news from Correspondent Author (23 October 2022)

1. Original Manuscript (5 October 2021)

A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction

Gregorius Satia Budhi*

School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia, Gregorius.Satiabudhi@uon.edu.au

Informatics Department, Petra Christian University, Surabaya 60236, Indonesia, Greg@petra.ac.id

Raymond Chiong*

School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia, Raymond.Chiong@newcastle.edu.au

Ensemble classifiers have been shown to perform well for a broad range of applications. The main idea behind these classifiers is to combine multiple standard single classifiers to produce a combination that outperforms any single classifier itself. In this study, we propose a novel ensemble model—the Multi-type Classifier Ensemble (MtCE). Unlike other ensemble models that utilise only one type of single classifier, our proposed ensemble utilises several different types of machine learning classifiers (including deep learning models) as its base classifiers. From the experiments, we find that the MtCE adequately detects fake reviews, which have become an acute problem on e-commerce platforms. The MtCE outperforms other ensemble methods on accuracy and other measurements in all the relevant public datasets we used. Moreover, if set correctly, the parameters of MtCE, such as base-classifier types, total number of base classifiers, bootstrap and the method via which to vote on output (e.g. majority or priority), further improve the performance of the proposed ensemble.

Keywords: Ensemble classifiers, machine learning, deep learning, fake review detection.

1 INTRODUCTION

Ensemble classifiers are increasingly used in many different areas of study, such as text analysis [1, 2], computer security [3, 4], environmental science [5], medical research [6], electrical fault detection [7] and spatial data processing [8]. The main idea behind ensemble classifiers is to combine multiple standard (single) classifiers to produce a combination that outperforms any single classifier itself [9, 10]. In general, ensemble classifiers can be classified into four groups: bagging, boosting, stacked generalisation, and the mixture of experts [11]. Bagging, also called bootstrap aggregating, accumulates multiple predictors/classifiers, and runs a plurality vote when predicting classes. These base classifiers are differentiated using a bootstrap technique to replicate the learning set [12]. The Random Forest (RF) and Pasting Small Votes ensemble models are a variation of bagging [13, 14]. The second group of ensemble models is boosting. Similar to bagging, boosting is also an aggregation of multiple classifiers. However, instead of using a bootstrap, it uses the boosting technique to train the base classifiers. This technique can convert weak learning algorithms to achieve arbitrarily high accuracy [15]. Adaboost (AB) [16] and Gradient-Boosted Trees (GB) [17] are two well-known algorithms that use this technique. In stacked generalisation [18], the classifiers are stacked to one another so that the output of some first-level base classifiers becomes the input of a second-level meta classifier. The mixture of experts method [19] combines the weighted output of several base classifiers in the consecutive combiner.

While promising, we still see potential for improvement from these groups of ensemble classifiers; they typically use only one type of single classifier as the base classifier, and differentiate each base classifier by bootstrapping or boosting their inputs via bagging or boosting, stacking the classifiers, or applying different weights for their output. In

* Corresponding authors' emails:
Raymond.Chiong@newcastle.edu.au, Greg@petra.ac.id, Gregorius.Satiabudhi@uon.edu.au

this study, we propose a novel ensemble model—the Multi-type Classifier Ensemble (MtCE)—that utilises several different types of single classifiers as its base. As every single classifier has strengths and weaknesses in classification or detection, by combining different types of single classifiers in an ensemble, we use the strength of one classifier to 'cover' for the weakness of the other classifiers, and vice versa. To further increase the performance of our model, we implement the bootstrap technique [12], and the method to vote on output (majority or priority) to produce the final output.

In this study, we focus on experimenting with and testing our model for fake review detection, especially via textual-based features. Fake review detection is a hot research topic in natural language processing [20]. A fake review is a consumer review with fictitious opinions written deliberately to sound authentic, for commercial motives; that is, to promote or damage the reputation of the reviewed product [21, 22]. There is a significant possibility that fake reviews distort the actual evaluation of a product [23], create harmful effects and eventually reduce trust in consumer reviews and undermine the effectiveness of the online market [24]. These fraudulent acts occur in response to the vital role of consumer reviews in purchasing behaviour in online markets [25]; consumers trust opinions expressed in online reviews and depend on them to make decisions [2, 26, 27]. Without detection efforts, reviews may be replete with lies, fakes and deceptions, and thus completely useless—or worse [20].

The majority of studies in fake review detection follow three main approaches: that based on the content of the reviews, that based on the behaviour of the reviewers, or a combination of these [28, 29]. Content-based approaches focus on the linguistic features of text such as words, parts of speech (POS), n-gram, term frequency (TF) or other linguistic characteristics [20, 30, 31]. The behaviour-based approach focuses more on reviewers' behaviour, such as the user's identity, reviewed product, total number of reviews, ratings given, and duration of reviews [28, 32, 33]. This approach is straightforward, needs only a small number of features, and can deliver good results if properly designed. However, the behaviour-based approach is entirely dependent on additional information, such as metadata, provided by the system. In contrast, the content-based approach is independent of the system, requiring only the review text.

The content-based approach has two sub-categories. The first creates input features from the review text using Bag of Words (BOW), Word2Vec and skip-gram; thus, the input features for detection are words or terms created from the review text itself, and we call this approach textual-based featuring. The features extracted by this approach are more or less similar, mainly in the form of n-gram terms [21, 26, 30, 31, 34]. This textual-based approach is promising and is used in many studies to extract features from review texts. While independent of the system and needing only the text, this approach usually produces good results [20, 21, 30, 31, 34-37]. The second form of content-based methods attempts to extract information and properties from the text, such as the length of the text, total words and total sentences, in addition to linguistic characteristics of the text, such as POS, subjectivity, complexity, diversity and similarity [38-43]. This second type of content-based featuring is lightweight compared with textual-based featuring; for example, only 80 features in [44] compared with 5,000 features in [45]. This approach, too, is independent of the system since it requires only the text; however, performance is usually relatively poor if not combined with other approaches [44].

To test the performance of the proposed MtCE, we used a textual-based approach, which is relatively independent of the system. For the experiments, we used Ott et al.'s [35] and Li et al.'s [37] datasets. These datasets are public fake review datasets used in many studies (e.g., see [20, 21, 30, 31, 36, 38, 44]). Before extracting the input features, we implemented text preprocessing such as tokenisation, stopword removal, detection of negation words, correction of elongation words, and POS lemmatisation. To extract the input features, we used the BOW method, n-gram words (from unigram to trigram words) and the count vectorised method. These methods convert a collection of text documents to a matrix of token counts. The token count features are in descending order by TF across the corpus (dataset). For the base classifiers of our proposed ensemble model, the MtCE, we implemented several standard single classifiers that performed well in previous research [1, 46], which include the Logistic Regression (LR), Linear-kernel Support Vector Machine (LSVM) [47], Multi-Layer Perceptron (MLP) [48], Decision Tree (DT) [49], Naïve Bayes (NB), and Convolutional Neural Network (CNN). For the comparison, we also applied other ensemble models such as the Bagging Predictors (BP) [12], RF, AB and GB.

The rest of this paper is organised as follows. In the next section, we explain the design of our proposed model, the MtCE. After that, we discuss how we conducted experiments to test the performance of the MtCE, such as the datasets we used, testing framework, details of the textual-based approach, types of base classifiers, and the measurements. In Section 4, we discuss the results and analyses, and finally, we draw conclusions and outline the possibilities for future work.

2 THE MTCE

The MtCE is a novel ensemble of classifiers developed in light of the fact that every single classifier has its strengths and weaknesses. In Figure 1, we depict two processes: the training process (the full head arrows ➜) and the testing process (the line head arrows →). All processes begin with inputting the training or testing sets and the parameter settings of the MtCE (see hollow head arrows -▷). For a training process, the first process determines if the base classifiers will be created in equal numbers or in random fashion. For example, suppose the total base classifiers are 10 of 3 types (LR, DT, NB); an equal setting will create 4 LR, 3 DT and 3 NB, while for a random setting, each classifier type will be randomised from 1 to 8 (total classifiers – (total types – 1)) classifiers. After creating the base classifiers, each base classifier is assigned a subset of training samples based on the bootstrap sampling process. The bootstrap sampling method draws sample data repeatedly with replacement from the source, based on the percentage setting [12]. After the training process for each base classifier finishes, the weights and parameters of all base classifiers are saved in a file.

The testing process is straightforward. After inputting the testing samples and loading the previously trained base classifiers, all samples are run with the base classifiers. The final output will be determined by one of two processes: the majority vote or the priority threshold. The majority vote chooses the majority class outputs from the results of the base classifiers. If the majority votes of all classes are equal, the final result is the smallest class in the corpus. The priority threshold provides a final result based on the priority chosen in the setting; that is, if the positive class is chosen, then, if the positive class outputs from base classifiers are the same as or more than the threshold, the final output is a positive class. The priority threshold can be set from absolute, which needs only one base classifier to pick the chosen class, to the maximum threshold before it becomes a majority vote (i.e. more than 50% + 1 in a binary classification problem).

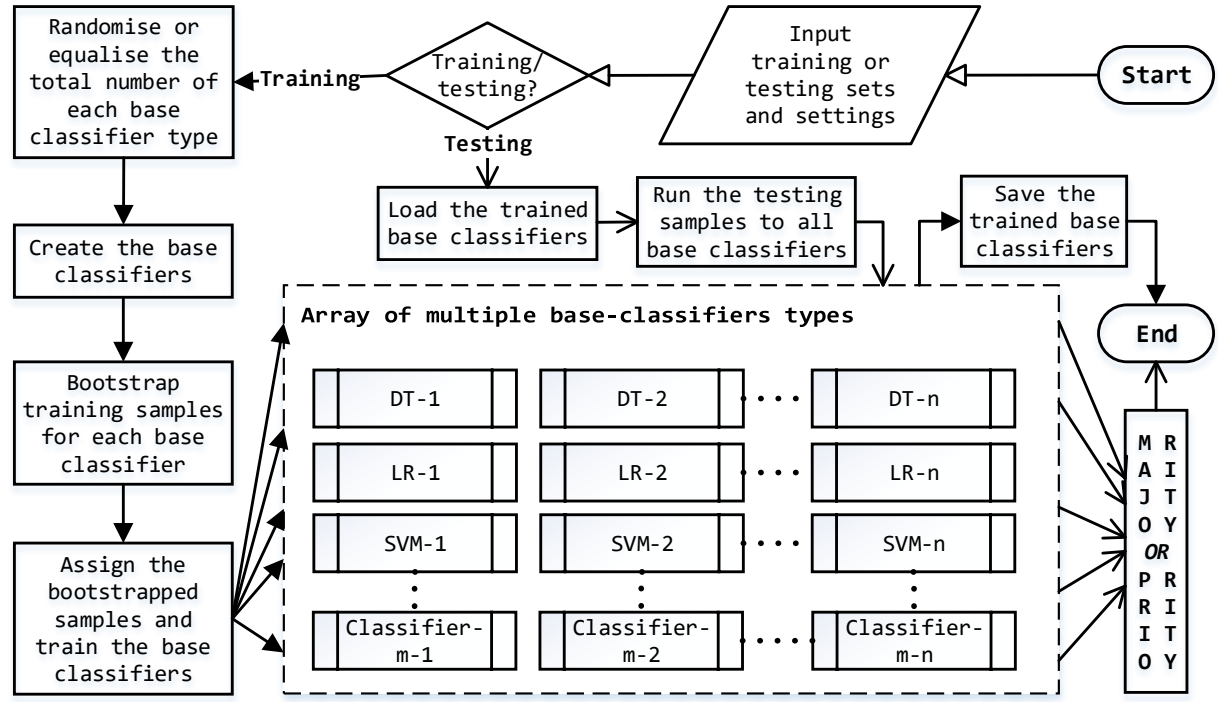

Figure 1: Design of the MtCE

3 TESTING THE PERFORMANCE OF THE MTCE

3.1 DATASETS

Intuitively, our proposed ensemble classifier can be applied to a wide range of problems, similar to other machine-learning ensemble classifiers. However, to test the performance of the MtCE, we conducted experiments using four public fake review datasets (see Table 1 for additional details). The public datasets from Ott et al. [33] and Li et al. [35] were created using domain experts, such as Amazon's Mechanical Turk service (Turkers) and hotel employees, to provide false/fake reviews for some online services. Li et al.'s hotel dataset is an extension of Ott et al.'s dataset.

Table 1: Statistics for several public datasets

Dataset Author	Domain	Total Records	Fake		Genuine	
			Total	%	Total	%
Ott et al. [35]	Hotel (Various)	1600	800	50.00	800	50.00
Li et al. [37]	Doctor (Various)	558	357	63.98	201	36.02
	Hotel (Various)	1880	1080	57.45	800	42.55
	Restaurant (Various)	402	202	50.25	200	49.75

3.2 FRAMEWORK FOR TESTING

To evaluate our proposed ensemble, the MtCE, we implemented the comparison framework that we used previously to investigate classifiers for sentiment polarity detection of customer reviews [1]. In this study, we modified the framework so that it could be applied to test the MtCE for fake review detection (see Figure 2). We used this framework

to test different settings of the MtCE using similar datasets and comparing their performances. We ran the same test using several single classifiers used as base classifiers of the MtCE and several well-known ensembles for comparison.

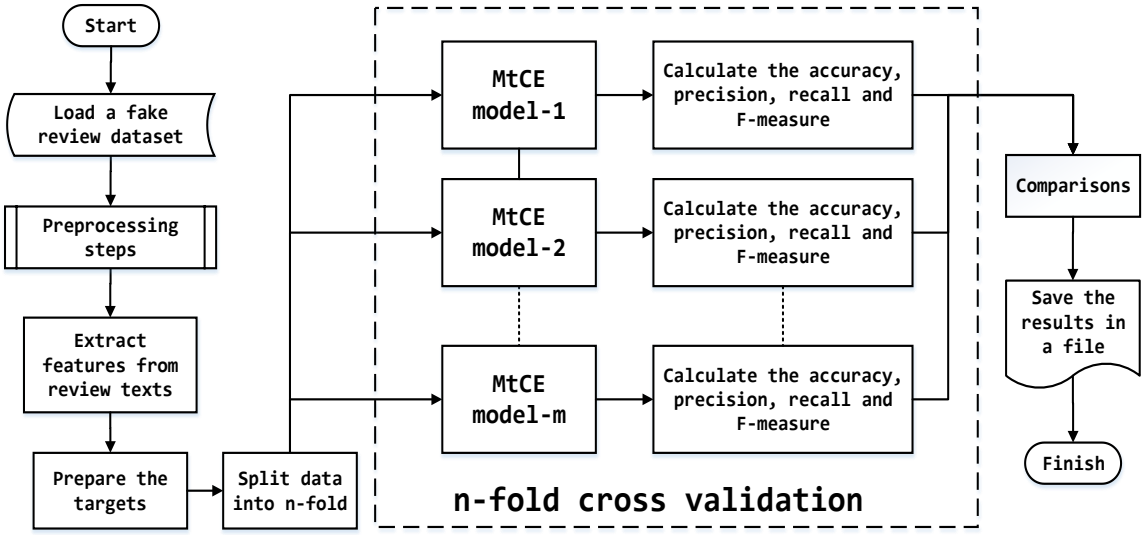

Figure 2: Framework to test and compare the MtCE

For the preprocessing of the text reviews (see Figure 3), we implemented several methods that performed well in previous studies [45], as follows:

a. removal of punctuation, numbers and stopwords
b. correction of spelling errors, elongation words and negative words
c. POS tagging and lemmatisation.

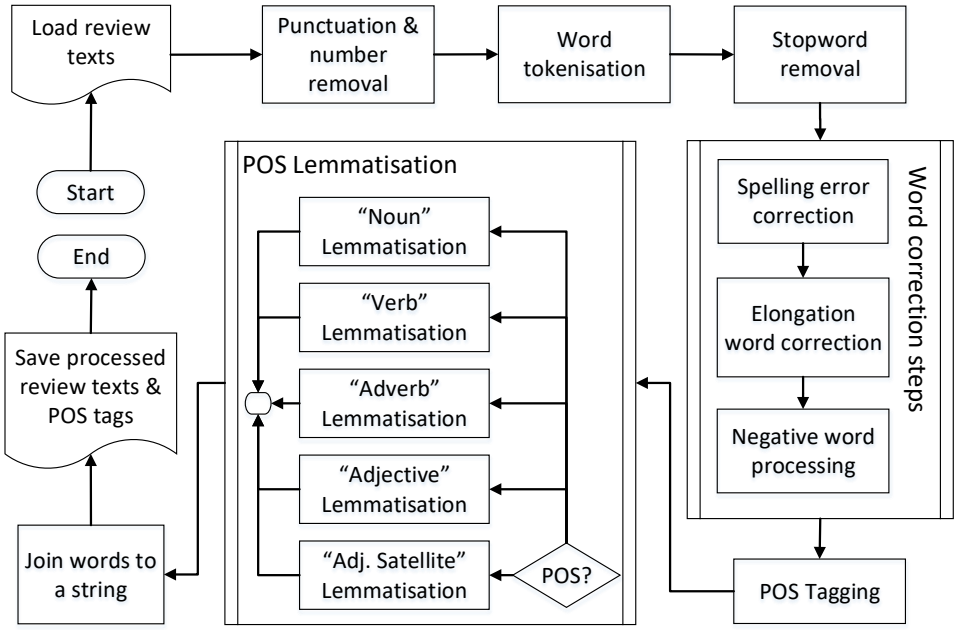

Figure 3: Preprocessing steps [45]

Punctuation and number removal is a necessary step for a textual-based approach since the input features of this approach are the words. Therefore, we needed to remove the non-word parts of the text reviews. Stopword is a term for words commonly used in English sentences such as ‘the’, ‘is’, ‘at’, ‘which’ and ‘on’. These words can be removed from the text before it is used as a training record. These kinds of words may mislead detectors in recognising the trait words of a class because they are often the most common words in a corpus.

Misspelled words create unnecessary issues for detection since the detector will recognise them as different words from their correct forms. Like misspelled words, elongation words such as ‘yesss’, ‘fiiine’ and ‘yoouu’ can also increase the diversity of words in the training example and make the classifiers harder to train. We utilised Peter Norvig’s code for spelling correction [50] to correct misspelled and elongation words. Negative words have many forms, depending on the grammar, but their purpose is similar—to change a positive sentence to be negative. Therefore, we shifted all negative words to their primary form to reduce the diversity of words.

POS tagging categorises words by their syntactic function, such as noun, pronoun, adjective, verb and adverb. This step is essential because it puts the words in their context [31] so that in the lemmatisation process, we can lemmatise the word to the correct context. Lemmatisation is a method for changing the word back to its basic form; POS lemmatisation returns the chosen word to its basic syntactic function. This step reduces the diversity of words inside the dataset and makes it easier to recognise.

Because we extracted the input features directly from social media messages, we used the BOW feature extraction method commonly used for textual-based features [51]. This method works by decomposing the entire text into a group of singular words. In this study, we used it in combination with the n-gram technique to capture singular words and terms that convey one meaning but are composed of multiple words. For terms, the BOW checks for the existence of a contiguous sequence of n words from the given text sample. This study limited this checking to trigrams since sequences of more than three words are rare in real-world texts. After decomposing the text, the terms were sorted based on their frequency across the corpus.

3.3 BASE CLASSIFIERS AND ENSEMBLE CLASSIFIERS

As mentioned in Section 2, our proposed ensemble can apply multiple types of single classifiers. While in theory, any single classifier could be used as the base classifier of the MtCE, to test its performance, we investigate several combinations of machine learning and deep learning that performed well in our previous study [1] as follows:

1. The LR classifier [52] is a member of the set of generalised linear models developed by Nelder and Wedderburn [53] and improved by Hastie and Tibshirani [54]. Linear models have limitations such as dependent variables that are continuous and normally distributed, which is not always desirable. Generalised linear models overcome this problem by using non-normal dependent variables [55, 56]. In LR, the dependent variables can either be unordered or ordered polytomous, while the independent predictor variables can either be interval/ratio or dummy variables [52].
2. The SVM is a supervised learning classifier that learns from training data and performs classification on new data. It separates different classes by a hyperplane and then maximises the separation distance to the greatest extent possible. The larger the margin, the lower the error generated by the classifier [47]. The linear kernel is generally recommended for text classification, so we used this with the SVM in our study (LSVM) [57].
3. The MLP is a feed-forward artificial neural network that uses supervised learning. This algorithm continually computes and updates all the weights in its network to minimise error. It consists of two phases: a feed-forward phase, where the training data are forwarded to the output layer, and a second phase. The difference between this output and the desired target (the error) is backpropagated to update the network weights [48]. In this study, we implemented the Adam optimiser to improve the performance of the algorithm [58].
4. The DT classifier is based on Hunt’s algorithm [59] as developed by Quinlan [49]. This algorithm builds a tree-like decision model for classification and prediction purposes and is a useful explanatory tool for expressing

the cause and effect chain [60]. DT is typically used as a base classifier for ensembles, such as BP, RF and AB.

5. The NB is often used in classification problems, including text classification [61, 62]. It is the simplest form of Bayesian network classifier if each feature is independent. Many applications have successfully implemented Naïve Bayes, which is considered one of the top 10 data mining algorithms [63]. In this study, we use Multinomial Naïve Bayes [64].
6. The CNN was invented by Lecun et al. [65] in 1998. In general, CNN consists of convolutional layers that create features for the network to learn. These convolutional layers can be complemented with normalisation layers and pooling layers. Typically, the convolutional layers are flattened with fully connected layers, followed by a softmax layer for classification or pattern recognition [66-68]. By varying the number of layers and nodes/neurons in each layer, CNN has fewer connections and parameters and is easy to train. Theoretically, its training performance is slightly poorer than the standard feed-forward neural network [66].

For comparison with the performance of the MtCE, we also ran several well-known ensemble classifiers on the same framework, as follows:

1. The RF is an ensemble of DT predictors where each tree is independently trained using a random vector. Error generalisation of an RF depends on the strength of each individual tree and the correlation between them. This ensemble model is relatively robust to outliers and noise [14].
2. The AB [16] ensemble algorithm iteratively combines multiple weak classifiers such as DT over several rounds. It starts with equal weights for all training data. When the training data points are misclassified, the weights of these data points are boosted and a new classifier is created using the new unequal weights. This process is repeated for a set of classifiers [69].
3. The BP ensemble model uses several single predictors to build a cluster of predictors. The predictors are trained in a bootstrapping process that replicates the training set. Plurality voting is utilised to predict a class [12]. By default, scikit-learn's BP, which we implemented, uses a DT as its base predictor.
4. The GB is an ensemble of gradient-boosted regression trees for classifying dirty data that produces a robust, competitive and interpretable algorithm for classification and regression. However, it uses only a single regression tree for binary classification [17].

We built all machine-learning classifiers and ensembles using scikit-learn components [70], except the CNN with Keras component [71] wrapped with scikit-learn component [70]. Scikit-learn does not have CNN component, but provide a way to wrap DL components of Keras to be used with or within scikit-learn. To ensure the results could only be affected by our model implementation, and not by modifying classifier parameters, we used the default parameters for all single classifiers used as base classifiers and the ensemble models.

3.4 CROSS-VALIDATION AND MEASUREMENTS

We evaluated the performance of the MtCE using n-fold Cross-Validation (CV), which is widely applied to evaluate the generalisation performance of an algorithm [4]. CV is mainly used in classification and regression models to reduce bias between the entire dataset and the training or the testing set [72]. Another purpose of using CV is to avoid overfitting [73]. Data are split into n disjoint folds or partitions in CV; n − 1 folds (subsamples) are used for training, and one is used as the testing sample. Therefore, when we consider n = 10, the process is repeated 10 times, and average measurements are calculated. We implemented scikit-learn [70] for performance measurements of our experiments. These measurements and their respective formulas can be found in Table 2.

Table 2: Measurement functions and formulas

No	Name	Sklearn Function	Equation

1	Accuracy	accuracy_score()	$A(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{k=0}^{n_{samples}-1} 1(\widehat{y_k} = y_k)$ where y is the set of predicted pairs, $\hat{y}$ is the set of true pairs and $n_{samples}$ is total samples.
2	Precision	precision_score()	$P(y_k, \widehat{y_k}) = \frac{tp}{tp + fp}$ where k is the set of classes, yk is the subset of y with class k, tp is true positive, and fp is false positive.
3	Recall	recall_score()	$R(y_k, \widehat{y_k}) = \frac{tp}{tp + fn}$ where fn is false negative.
4	F-measure/F1	f1_score()	$F_1(y_k, \widehat{y_k}) = 2 * \frac{P(y_k, \widehat{y_k}) * R(y_k, \widehat{y_k})}{P(y_k, \widehat{y_k}) + R(y_k, \widehat{y_k})}$

4 RESULTS AND DISCUSSION

4.1 EXPERIMENTING WITH THE MTCE FOR FAKE REVIEW DETECTION

The first group of experiments aimed to test the performance of the MtCE for fake review detection. For the base classifiers, we implemented CNN, LR, LSVM, MLP, DT and NB single classifiers. These base classifier parameters are the defaults of the scikit-learn components used to implement these classifiers [70]. As shown in the Appendix, we tested all possibilities from 2-, 3-, 4- to 5-combination and presented 11 selected combinations in Table 3. Besides combining base classifiers, the other settings were fixed: total base classifiers = 15, shared equally for each type; bootstrap = 0.5; and final output = majority vote. The datasets we used in the experiments were Ott et al.'s dataset [35] and all three of Li et al.'s datasets [37]. All experiments were conducted using the 10-fold CV method. For the comparison, we also predicted the datasets using single classifiers as above and several well-known ensemble classifiers (RF, BP, AB and GB). The results of the single and ensemble classifiers are provided in Table 4.

Table 3: Results of the combination of base classifiers for the MtCE model (selected)

Num[1]	MtCE Base Classifier Combination	Ott et al.'s Dataset[2]				Li et al.'s Dataset (Doctor)[2]			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
2	LR-MLP	**88.00**	*86.85*	**89.44**	**88.08**	84.41	*84.23*	**92.97**	88.33
5	LSVM-MLP	87.50	85.72	**90.02**	**87.74**	**86.56**	**87.05**	**93.12**	**89.88**
7	LSVM-NB	**89.19**	**88.32**	**90.43**	**89.31**	*84.75*	86.85	**90.32**	88.19
9	MLP-NB	**89.63**	**88.52**	**91.22**	**89.80**	*85.31*	*84.00*	**95.30**	*89.12*
18	LR-LSVM-NB	**89.00**	**88.68**	**89.59**	**89.07**	84.58	85.70	**91.44**	88.34
22	LSVM-MLP-NB	**89.13**	87.86	**91.00**	**89.31**	85.13	86.39	91.88	88.79
26	CNN-LR-MLP	**88.19**	**87.32**	**89.33**	**88.27**	84.40	83.75	**93.51**	88.29
35	LR-LSVM-MLP-NB	**89.19**	**88.77**	**89.95**	**89.28**	85.66	87.01	91.38	88.94
37	LR-MLP-DT-NB	**88.75**	**88.41**	89.25	**88.79**	84.06	83.85	**93.06**	88.04
47	LR-LSVM-MLP-DT-NB	**88.94**	**88.38**	**89.56**	**88.94**	84.24	84.89	92.00	88.12
50	CNN-LR-LSVM-MLP-NB	**88.88**	**88.21**	**89.83**	**88.95**	86.21	86.45	**93.18**	89.54
		Li et al.'s Dataset (Hotel)[2]				Li et al.'s Dataset (Restaurant)[2]			
2	LR-MLP	**87.34**	85.09	**94.30**	**89.45**	85.11	83.99	**88.36**	85.69

5	LSVM-MLP	86.12	84.71	**92.67**	**88.44**	**85.83**	83.98	**88.55**	**86.02**
7	LSVM-NB	**87.45**	87.09	**91.75**	**89.34**	85.35	84.51	86.33	85.21
9	MLP-NB	87.23	86.69	**91.92**	**89.17**	86.04	85.36	87.18	**86.15**
18	LR-LSVM-NB	85.96	85.20	**91.34**	88.14	**86.58**	85.71	**88.08**	**86.41**
22	LSVM-MLP-NB	86.97	86.09	**92.34**	**89.05**	85.32	83.68	**88.65**	85.75
26	CNN-LR-MLP	85.85	84.79	**91.96**	88.17	83.34	82.42	84.09	82.88
35	LR-LSVM-MLP-NB	86.70	85.12	**92.92**	88.78	82.32	81.51	84.77	82.80
37	LR-MLP-DT-NB	**87.82**	86.67	**93.26**	**89.81**	84.38	83.61	85.67	84.57
47	LR-LSVM-MLP-DT-NB	**87.55**	86.19	**93.28**	**89.58**	84.63	83.29	87.36	85.06
50	CNN-LR-LSVM-MLP-NB	86.76	85.98	**91.98**	**88.81**	83.80	82.59	86.17	83.98

[1] The number here is associated with the number in Appendix
[2] **Bold-green** = the MtCE result is higher than that of all base classifiers; *Italic-red* = the MtCE result is higher than all base classifiers; Normal-black = at least one base classifier result is higher than that of the MtCE.

Table 4: Results of single and ensemble classifiers

Num.	Classifiers	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	CNN	79.63	81.48	77.36	79.13	67.76	70.64	88.40	77.02
2	LR	87.13	87.05	87.25	87.11	83.35	85.48	89.71	87.26
3	LSVM	85.88	85.67	86.21	85.89	83.13	86.28	87.80	86.91
4	MLP	87.56	87.05	88.37	87.66	85.67	85.89	92.94	89.16
5	DT	71.50	72.21	69.71	70.83	65.79	72.98	74.01	73.32
6	NB	88.50	87.97	89.45	88.63	87.12	90.68	88.98	89.68
7	RF	87.00	87.44	86.62	86.92	76.35	74.05	97.24	83.98
8	BP	75.00	74.23	76.79	75.41	74.19	74.50	90.21	81.45
9	AB	80.06	79.64	80.82	80.09	74.93	79.56	81.28	80.16
10	GB	82.81	83.67	81.63	82.56	76.52	77.08	90.63	82.99
		Li et al.'s Dataset (Hotel)				Li et al.'s Dataset (Restaurant)			
1	CNN	78.83	82.48	80.63	81.29	71.63	71.66	73.20	71.75
2	LR	85.59	85.50	90.25	87.77	81.59	81.03	84.52	81.74
3	LSVM	84.20	84.59	88.26	86.33	82.11	79.93	84.40	81.53
4	MLP	86.12	85.89	90.82	88.24	85.56	84.85	87.69	85.73
5	DT	68.88	73.81	71.15	72.38	60.24	60.34	63.27	61.11
6	NB	87.29	89.73	87.92	88.78	86.08	86.36	86.63	86.10
7	RF	81.97	80.78	90.28	85.17	81.34	78.31	87.86	81.88
8	BP	75.96	76.35	84.14	80.03	71.65	71.08	74.51	71.95
9	AB	79.89	80.77	84.87	82.67	70.89	70.64	70.35	70.38
10	GB	80.48	78.83	89.94	84.00	76.12	77.43	75.46	75.78

Comparing Table 3 with Table 4, some combinations of base classifiers in the MtCE exceed the performance of all of the base classifiers (see the bold-green scores in Table 3). For example, accuracy of the MtCE(LR-MLP) for Ott et al.'s = 88%; in comparison, in Table 4, accuracy of LR and MLP are 87.13% and 87.56%. However, not all combinations improved and supported each other as we hoped. A few weak classifiers degraded the performance of stronger

classifiers when combined with these. This gave rise to the result that the performance of the MtCE was in the middle of the performance of its base classifiers; for example, LSVM accuracy for Ott et al.'s = 85.88% while MLP = 87.56%, and the combination of both in MtCE(LSVM-MLP) = 87.50%. This implies that LSVM degraded the performance of MLP, and given this, rather than implementing our proposed ensemble, it would be better to use MLP directly. However, in the same case, MtCE(LSVM-MLP) for Ott et al.'s improved the recall, from 86.21% (LSVM) and 88.37% (MLP) to 90.02%. In a case such as this, using our ensemble would be acceptable since the recall in binary classification/detection is the same as the accuracy of the positive class or the main class (in our problem, the fake review class). Thus, high recall means the ability of the ensemble to detect positive classes is also high.

Another finding from this set of experiments is that different datasets might need different base classifier combinations. For example, the MtCE(MLP-NB) that performed best for Ott et al.'s dataset performed worst for Li et al.'s doctor dataset (increasing the recall but decreasing all other measurements). The best combinations for Li et al.'s doctor dataset was MtCE(LSVM-MLP); for Li et al.'s hotel dataset was MtCE(LR-MLP-DT-NB); and for Li et al.'s restaurant dataset was MtCE(LR-LSVM-NB).

We found that the MtCE performed better than other ensembles of classifiers, as indicated by comparing the results in Table 4 (numbers 7–10) to the MtCE results in Table 3. These facts suggest that, when performed correctly, the combination of multi-type classifiers could cover the weakness of each and outperform the combination of a homogenous type of classifier such as RF, BP, AB or GB. However, for the cases of RF, BP and AB, we suspect the cause is also because these ensembles use DT as their base classifiers/predictors. DT's performance was not good on all datasets, which therefore affected the performance of its ensemble variants.

While deep-learning CNN as a stand-alone classifier did not perform as well as other single classifiers, somewhat surprisingly, it performed exceptionally well when combined with other classifiers in the MtCE. This increase in accuracy was significant; for example, MtCE(CNN-LR-LSVM-MLP-NB) is 7.93% in Li et al.'s hotel dataset to 18.45% in Li et al.'s doctor dataset. The improvement in recall was also good, with a minimum of 4.78% in Li et al.'s doctor dataset to 12.97% in Li et al.'s restaurant dataset. Therefore, the deep-learning CNN could be combined perfectly with other machine-learning single classifiers as base classifiers for the MtCE; it gave good results and improved all other base classifiers' performances as well.

4.2 EFFECTS OF MTCE PARAMETERS

In this section, we discuss several input parameters of the MtCE that may affect the performance of this proposed model. All of the settings are the same as in Section 4.1 except the parameter we are testing. For these experiments, we chose four base classifier combinations of the MtCE in Table 3—that is, MtCE(LSVM-MLP), MtCE(LSVM-NB), MtCE(LR-LSVM-NB), MtCE(LR-MLP-DT-NB) and MtCE(LR-LSVM-MLP-DT-NB)—as these are the best five combinations across all four datasets. We ran all experiments for the parameter testing using Ott et al.'s dataset only, since these five MtCE combination settings perform well in this dataset (see the bold-green scores in Table 3). However, after finding the ideal set of parameters, we ran them with the other datasets and compared the results with our previous approach implemented on the same datasets.

First, we investigated the effect of total base classifiers. Here, we ran experiments on all five combinations with total number of base classifiers varying from 5 to 100. Results can be seen in Figure 4 for accuracy, Figure 5 for precision, Figure 6 for recall and Figure 7 for F-measure. From these figures, we can see that accuracy, precision, recall and F-measure were increasing at first, then, after 20 classifiers, all scores fluctuated around a number ± 1.5%. Accuracy fluctuated at 88.5%, precision at 88% and recall at 90%, except for MtCE(LSVM-MLP), which was 1% lower. However, we used the best accuracy results for each MtCE combination for the next batch of experiments. These combinations are 20 classifiers for MtCE(LSVM-MLP), 55 classifiers for MtCE(LR-LSVM-NB), 80 classifiers for MtCE(LR-MLP-DT-NB) and 85 classifiers for MtCE(LR-LSVM-MLP-DT-NB).

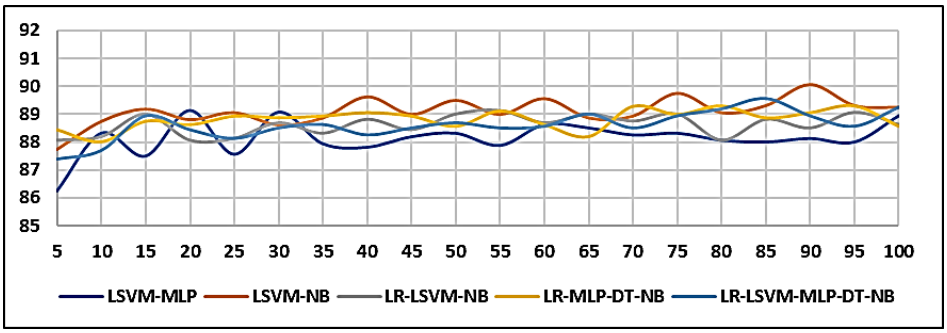

Figure 4: The accuracy of the MtCE with 5–100 total base classifiers

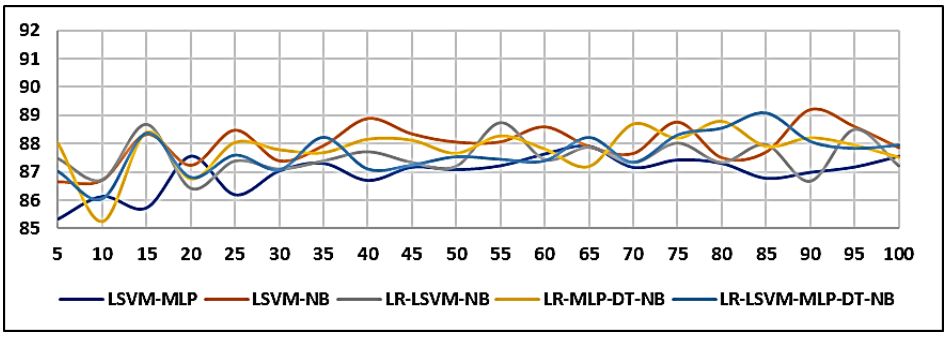

Figure 5: The precision of the MtCE with 5–100 total base classifiers

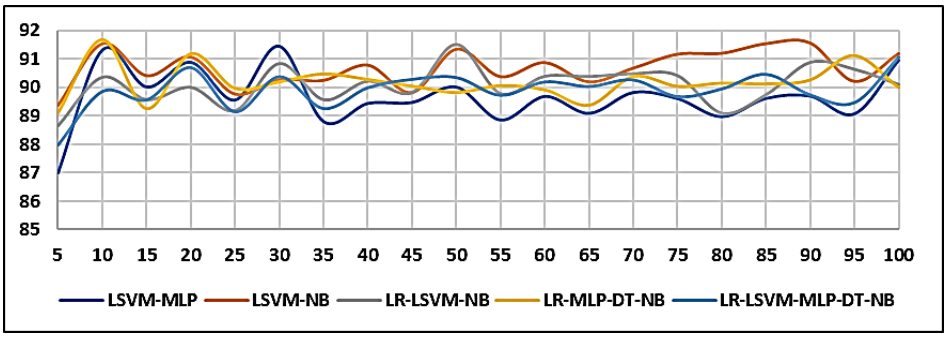

Figure 6: The recall of the MtCE with 5–100 total base classifiers

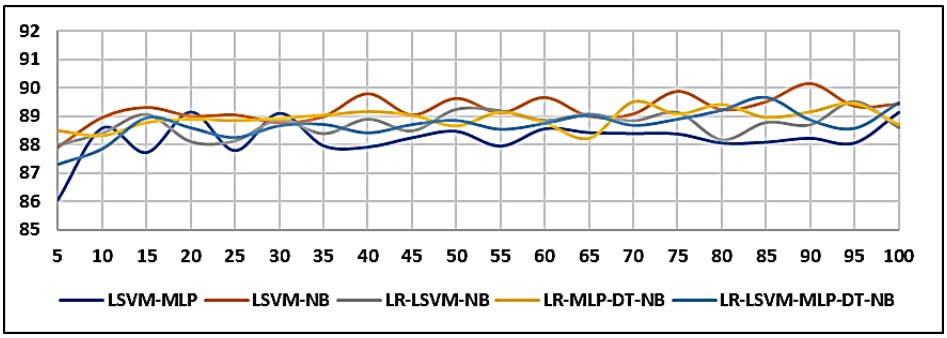

Figure 7: The F-measure of MtCE with 5–100 total base classifiers

The subsequent experiments aimed to investigate the bootstrap parameter. We used fixed parameters as above, except the bootstrap setting ranged from 0.25 to 1 (no bootstrapping). As can be seen in Figure 8, the best accuracy,

precision and recall were often achieved when the bootstrap setting was 0.5. Every base classifier was trained using 50% of training samples randomly. For the next set of experiments, we set the bootstrap to be 0.5.

To investigate the equally split vs randomly assigned base classifier type, we ran the experiments 10 times for each equally split and randomly assigned setting for each MtCE combination. The average results and their standard deviation can be seen in Table 5. This table shows that the equally split option gave slightly better average results for all measurements relative to the random option. This implies that when the total of each base classifier is equal, they support each other, and the strength of one type of classifier covers the weakness of the other type when combined in the MtCE. On the other hand, the random option could make one or two base classifiers more numerous than the others, so that they dominated the detection process. The standard deviation of all measurement scores below 1 means the variation of the results from 10 runs was low and close to each other. In more detail, we can see that the standard deviation of 2-combination, the MtCE(LSVM-MLP), on the equally split setting was lower than the random assignment. This is expected since, in an equally split setting, each base classifier total is the same for all 10 runs, while in the random option, this varies. However, the exact opposite happened in 5-combination, the MtCE(LR-LSVM-MLP-DT-NB). Here, the standard deviation of the equally split setting was higher than the random option, implying that the randomness might make the base classifiers more homogenous than split equally between its 5 base classifier types. For the 3- and 4-combinations, the standard deviation scores are similar.

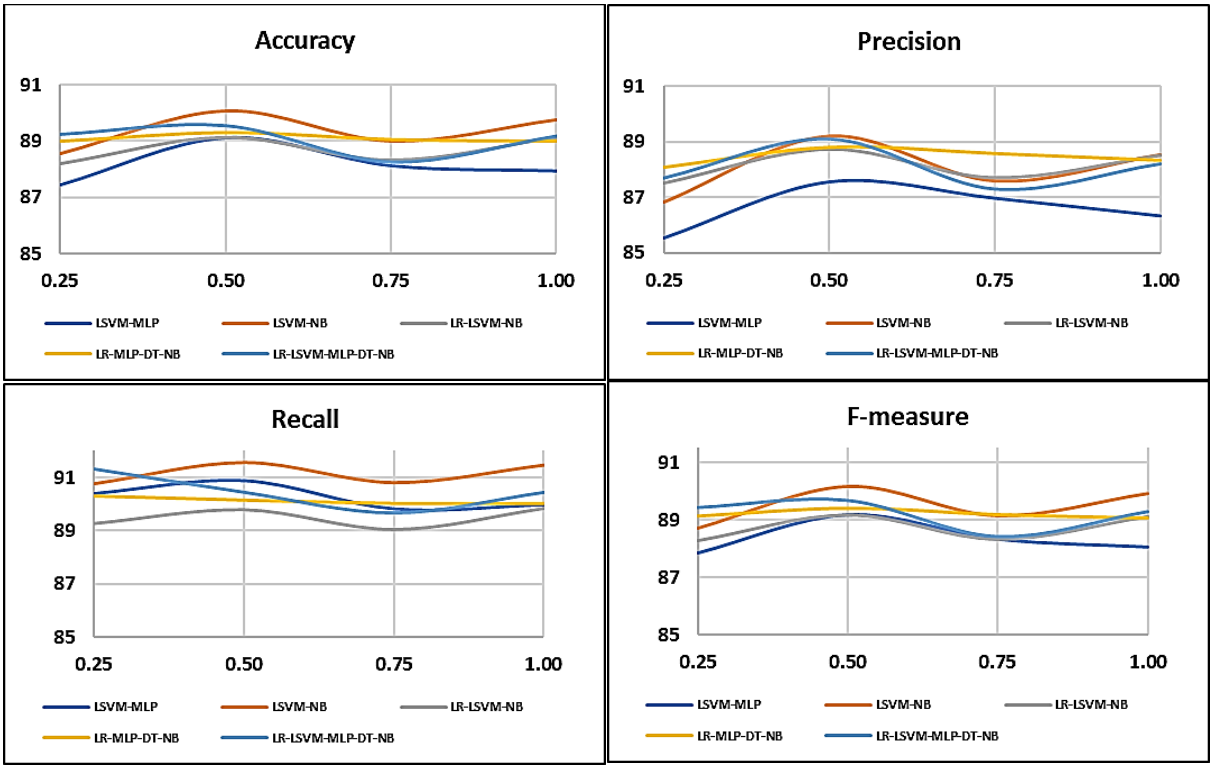

Figure 8: The accuracy, precision, recall and F-measure of the MtCE with bootstrap setting from 0.25 to 1.00.

Table 5. Results of equally split vs randomly assigned type of MtCE base classifiers on 10 runs of the 10-fold CV process.

Base Classifier Combination	Equally or Randomly Assigned	Accuracy		Precision		Recall		F-measure	
		Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev
LSVM-MLP	Equal	88.21	0.38	86.76	0.35	90.23	0.56	88.39	0.41
	Random	87.82	0.60	86.41	0.59	89.81	0.69	88.00	0.60

LSVM- NB	Equal	89.44	0.36	88.20	0.50	91.12	0.39	89.57	0.37
	Random	89.37	0.19	88.16	0.38	90.93	0.35	89.47	0.22
LR-LSVM-NB	Equal	88.97	0.35	88.14	0.48	90.06	0.42	89.04	0.35
	Random	88.72	0.37	87.79	0.41	89.89	0.49	88.78	0.38
LR-MLP-DT-NB	Equal	89.18	0.32	88.34	0.44	90.24	0.39	89.21	0.29
	Random	89.03	0.35	88.19	0.46	90.12	0.38	89.09	0.35
LR-LSVM-MLP-DT-NB	Equal	88.28	0.66	87.37	0.64	89.59	0.86	88.41	0.67
	Random	88.12	0.37	87.10	0.39	89.49	0.49	88.21	0.39

The last parameter to test is how the output of base classifiers should be processed. There are two options in terms of transferring the output of base classifiers to the final output: majority vote and priority class. To test the effect of priority, we set the priority to the positive class (fake class), in the range between absolute priority and 45% priority. As mentioned above, absolute priority means the final result will be recorded as positive/fake if one or more of the base classifiers records this. Percentage priority means the final result will be recorded as positive/fake if the number of base classifiers with this result is the same or more than the percentage threshold. The results of the experiments can be seen in Figure 9.

As per Figure 9, priority setting increases the recall but decreases other measurements. Moreover, as noted, the higher the recall, the more accurate the positive class (fake class) detection. Thus, we can conclude that the priority setting is helpful when samples of positive class are scarce or when the focus of research is on anomaly detection. The key is how high we choose the percentage of priority so that the increase in recall does not greatly decrease the other measurements. For example, in Figure 9, for 25% priority of MtCE(LSVM-MLP), where recall increased by 5.05%, the accuracy, precision and F-measure decreased by 2.75%, 7.15% and 1.70%, respectively.

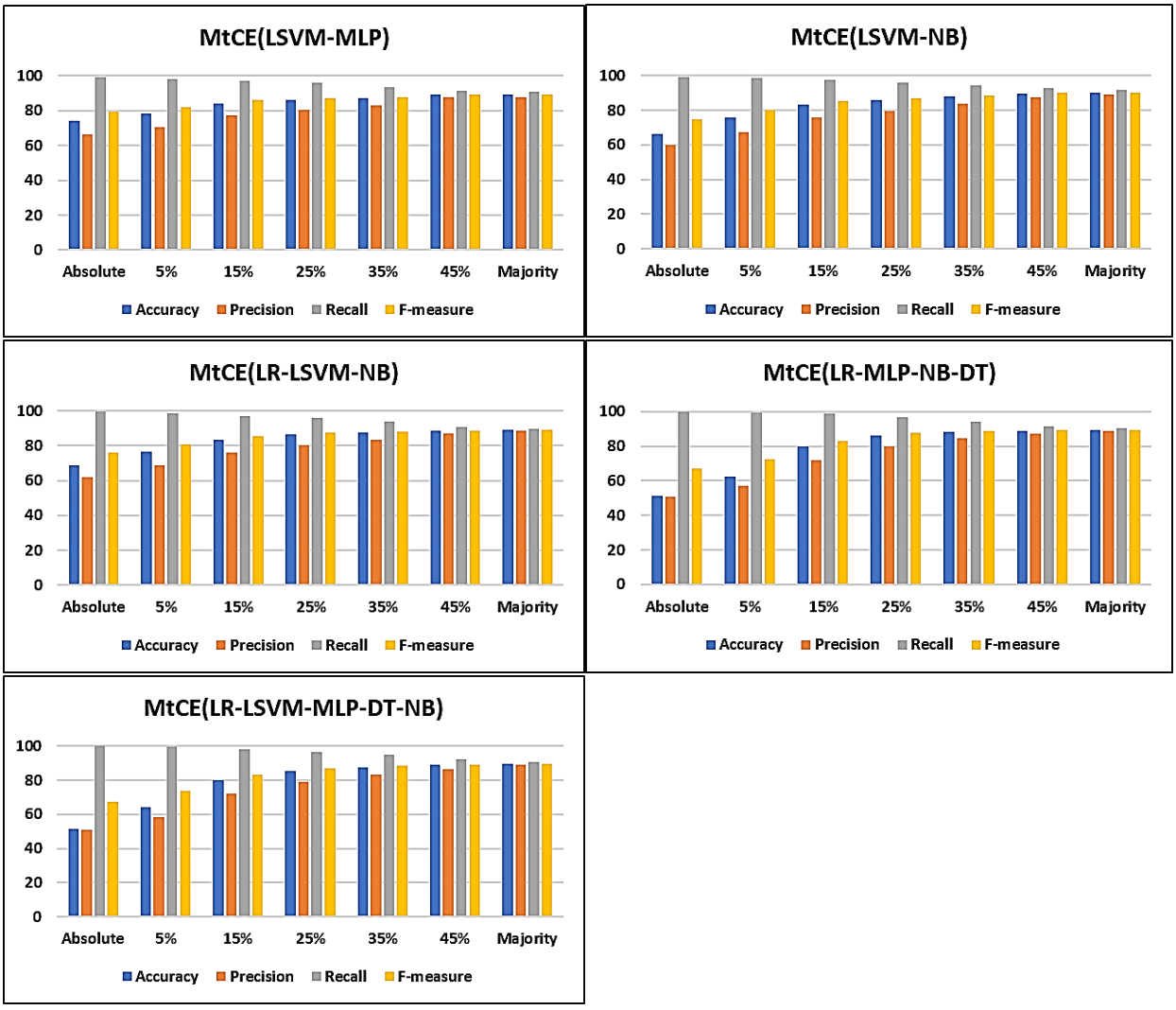

Figure 9: Results of final target experiments, from absolute priority to 45% priority, and majority.

4.3 COMPARISON TO OTHER STUDIES ON THE SAME DATASETS

In this section, we present the MtCE's and other models' best results in light of various considerations. Many factors can influence results, such as the measurement formulas or tools, datasets used, number of records involved in experiments, and how the experiments were conducted (e.g., CV vs traditional train-test-validation split). Therefore, the results presented in Table 6 should be read with the following considerations in mind:

1. We present the MtCE's and other studies' results based on the same datasets (see Table 1). We used all samples that could be processed from each dataset. Note that some studies used only a subset of the dataset or split the dataset into subsets and measured each subset differently. All of our experiments were conducted using 10-fold CV.
2. It is impossible to ensure that all the studies used the same formulas to measure accuracy, precision, recall and F-measure. Therefore, MtCE results were measured using widely used formulas for binary target prediction (see Table 2). All measurements were in the 'macro' average setting using scikit-learn measurement components [70].
3. The original results from the dataset creators were used as the baseline (see the underlined description and scores in Table 6. We present only the best result for each dataset from each study.

Table 6: Best performance of the MtCE and other studies for the same datasets

No	Dataset	Description	Acc (%)	Pre (%)	Rec (%)	F1 (%)
1	Hotel (various) [35]	Ott et al. [35], SVM on positive sentiment subset	88.4	89.1	87.5	88.3
		Fusilier et al. [30], PU-L modified	-	85.2	72.8	78
		Rout et al. [38], DT	92.11	-	-	-
		Etaiwi and Naymat [31], SVM, all preproc. steps	85	51	86	-
		Zhang et al. [36], DRI-RCNN, 0.9 T.P.	-	-	-	86.59
		Budhi et.al. [44], GB, over-sampling	70.62	67.58	81.29	73.8
		*MtCE(LSVM-NB, 90, Equal, 0.5, Majority)**	*90.06*	***89.19***	***91.56***	***90.16***
2	Hotel (various) [37]	Li et al. [37], OvR SAGE-Unigram	81.8	81.2	84	-
		Ren and Ji [20], Integrated, Employee/Turker subset	92.6	-	-	90.1
		Li et al. [21], SWNN	-	84.1	87	85
		Budhi et.al. [44], GB, under-sampling	71.29	73.61	82.79	77.93
		*MtCE(LR-MLP-DT-NB, 15, Equal, 0.5, Majority)**	*87.82*	***86.67***	***93.26***	*89.81*
3	Restaurant (various) [37]	Li et al. [37], OvR SAGE-Unigram	81.7	84.2	81.6	-
		Ren and Ji [20], Integrated, Customer/Turker subset	86.9	-	-	86.8
		Li et al. [10], SWNN		83.3	88.2	81
		Budhi et.al. [44], GB, over-sampling	72.44	71.23	75.16	73.14
		*MtCE(LR-MLP-DT-NB , 80, Equal, 0.5, Majority)**	*86.07*	***84.47***	***88.89***	*86.37*
4	Doctor (various) [37]	Li et al. [37], OvR SAGE-Unigram	74.5	77.2	70.1	-
		Ren and Ji [20], Integrated, Customer/Turker subset	76	-	-	74.1
		Li et al. [10], SWNN	-	83.7	87.6	82.9
		Budhi et.al. [44], GB, over-sampling	73.35	79.44	86.94	83.02
		*MtCE(LSVM-MLP, 15, Equal, 0.5, Majority)**	***86.56***	***87.05***	***93.12***	***89.88***

* MtCE(<type of base classifiers>, <total base classifiers>, <equally split or random assign of base classifiers>, <bootstrap percentage>, <majority or priority output>)

We do not claim that the MtCE is better or worse than methods in other studies since several factors can affect results. However, we can confidently compare the MtCE to our previous study in fake review detection [44] because the measurements are similar. As is evident in Table 6, compared with our previous approach, the MtCE is superior. Accuracy improves from a minimum of 13.2% in Li et al.'s doctor dataset to a maximum of 19.4% in Ott et al.'s dataset; precision improves from 7.6% to 21.6% and recall from 6.2% to 13.7%.

5 CONCLUSION AND FUTURE WORK

From the experiments above, we can conclude that our proposed ensemble, the MtCE, was able to adequately detect fake or fraudulent reviews, which have become an acute problem on e-commerce platforms. Compared with our previous approach, the MtCE improved accuracy up to 19.4%, precision up to 21.6% and recall up to 13.7%.

Not all combinations of base classifier types produced the expected result of an improvement in the performance on all base classifiers of the MtCE. In some combinations, weak classifiers dragged down the performance of stronger classifiers, especially in terms of accuracy measurements. However, when the combinations were correct, the MtCE produced better results than its base classifiers in stand-alone modes. It is important to note that while accuracy was not on top, recall was the highest in many cases. This is a positive result since, in binary classification, recall is the same as accuracy of a positive case. In this case, the MtCE was able to better detect the fake reviews class. Comparison with well-known ensembles, such as RF, AB, BP and GB, showed that some correct combination of the MtCE could outperform the other ensembles, proving that combining multi-type classifiers in one ensemble process performed better than combining the same classifiers, as is usually done in other ensemble methods.

We proved that deep-learning methods such as CNN could be combined with machine-learning counterparts as base classifiers to give better results than if run alone. The number of base classifiers affected performance detection; however, accuracy and other measurements oscillated after 25 base classifiers. While reducing the amount of data for the training process, the bootstrap setting could also affect the performance of the MtCE. From the experiments, we found that 50% bootstrapping gives better results than other percentages, including the non-bootstrap (bootstrap setting 100%). While the majority vote setting for the output offers a fair chance for each class to be detected, the priority threshold boosts detection of the prioritised class. This would be useful if the dataset is imbalanced or the MtCE is used to detect/classify anomalies.

Despite the extensive experimental studies presented in this paper, there remain opportunities/possibilities for further improvement. For our future work, we plan to investigate the implementation of the MtCE for multi-class problems, heavily imbalanced datasets and anomaly detection. The other avenue of research is optimising the parameters of the MtCE using our method to optimise ensemble parameters, the Multi-level Particle Swarm Optimisation [10].

REFERENCES

[1] Budhi, G. S., Chiong, R., Pranata, I. and Hu, Z. Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis. *Archives of Computational Methods in Engineering*, 28 (2021), 2543–2566.

[2] Kumar, N., Venugopal, D., Qiu, L. and Kumar, S. Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems*, 35, 1 (2018), 350-380.

[3] Hu, Z., Chiong, R., Pranata, I., Susilo, W. and Bao, Y. *Identifying malicious web domains using machine learning techniques with online credibility and performance data*. City, 2016.

[4] Hu, Z., Chiong, R., Pranata, I., Bao, Y. and Lin, Y. Malicious web domain identification using online credibility and performance data by considering the class imbalance issue. *Industrial Management & Data Systems*, 119, 3 (2019), 676-696.

[5] Shin, J., Yoon, S., Kim, Y., Kim, T., Go, B. and Cha, Y. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics*, 61 (2021).

[6] Xie, F., Fan, H., Li, Y., Jiang, Z., Meng, R. and Bovik, A. Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model. *IEEE Trans Med Imaging*, 36, 3 (Mar 2017), 849-858.

[7] Mohammadi, Y., Salarpour, A. and Chouhy Leborgne, R. Comprehensive strategy for classification of voltage sags source location using optimal feature selection applied to support vector machine and ensemble techniques. *International Journal of Electrical Power & Energy Systems*, 124 (2021).

[8] Jiang, Z., Sainju, A. M., Li, Y., Shekhar, S. and Knight, J. Spatial ensemble learning for heterogeneous geographic data with class ambiguity. *ACM Transactions on Intelligent Systems and Technology*, 10, 4 (2019), 1-25.

[9] Ren, Y., Zhang, L. and Suganthan, P. N. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. *IEEE Computational Intelligence Magazine*, 11, 1 (2016), 41-53.

[10] Budhi, G. S., Chiong, R. and Dhakal, S. Multi-level particle swarm optimisation and its parallel version for parameter optimisation of ensemble models: a case of sentiment polarity prediction. *Cluster Computing*, 23 (2020), 3371-3386.

[11] Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6, 3 (2006), 21-45.

[12] Breiman, L. Bagging predictors. *Machine Learning*, 24, 2 (1996), 123-140.

[13] Breiman, L. Pasting Small Votes for Classification in Large Databases and On-Line. *Machine Learning*, 36, 1 (1999/07/01 1999), 85-103.

[14] Breiman, L. Random forests. *Machine Learning*, 45, 1 (2001), 5-32.
[15] Schapire, R. E. The strength of weak learnability. *Machine Learning*, 5, 2 (1990/06/01 1990), 197-227.
[16] Freund, Y. and Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55, 1 (1997/08/01/ 1997), 119-139.
[17] Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 5 (2001), 1189-1232.
[18] Wolpert, D. H. Stacked generalization. *Neural Networks*, 5, 2 (1992/01/01/ 1992), 241-259.
[19] Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. Adaptive Mixtures of Local Experts. *Neural Computation*, 3, 1 (1991), 79-87.
[20] Ren, Y. and Ji, D. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385-386 (2017), 213-224.
[21] Li, L., Qin, B., Ren, W. and Liu, T. Document representation and feature combination for deceptive spam review detection. *Neurocomputing*, 254 (2017), 33-41.
[22] Ott, M., Choi, Y., Cardie, C. and Hancock, J. T. *Finding deceptive opinion spam by any stretch of the imagination*. City, 2011.
[23] Feng, V. W. and Hirst, G. *Detecting deceptive opinions with profile compatibility*. City, 2013.
[24] Malbon, J. Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, 36, 2 (2013), 139-157.
[25] Wu, Y., Ngai, E. W. T., Wu, P. and Wu, C. Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, 132 (2020).
[26] Cardoso, E. F., Silva, R. M. and Almeida, T. A. Towards automatic filtering of fake reviews. *Neurocomputing*, 309 (2018), 106-116.
[27] Bing, L., Wong, T.-L. and Lam, W. Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. *ACM Transactions on Internet Technology*, 16, 2 (2016), 1-17.
[28] Barbado, R., Araque, O. and Iglesias, C. A. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56, 4 (2019), 1234-1244.
[29] Heydari, A., Tavakoli, M. a., Salim, N. and Heydari, Z. Detection of review spam: A survey. *Expert Systems with Applications*, 42, 7 (2015), 3634-3642.
[30] Hernández Fusilier, D., Montes-y-Gómez, M., Rosso, P. and Guzmán Cabrera, R. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management*, 51, 4 (2015), 433-443.
[31] Etaiwi, W. and Naymat, G. The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science*, 113 (2017), 273-279.
[32] Savage, D., Zhang, X., Yu, X., Chou, P. and Wang, Q. Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications*, 42, 22 (2015), 8650-8657.
[33] Akram, A. U., Khan, H. U., Iqbal, S., Iqbal, T., Munir, E. U. and Shafi, M. Finding rotten eggs: A review spam detection model using diverse feature sets. *KSII Transactions on Internet and Information Systems*, 12, 10 (2018).
[34] Sun, C., Du, Q. and Tian, G. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering*, 2016 (2016), 1-7.
[35] Ott, M., Cardie, C. and Hancock, J. T. *Negative deceptive opinion spam*. City, 2013.
[36] Zhang, W., Du, Y., Yoshida, T. and Wang, Q. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management*, 54, 4 (2018), 576-592.
[37] Li, J., Ott, M., Cardie, C. and Hovy, E. *Towards a general rule for identifying deceptive opinion spam*. City, 2014.

[38] Rout, J. K., Singh, S., Jena, S. K. and Bakshi, S. Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications*, 76, 3 (2016), 3187-3211.
[39] Zhang, D., Zhou, L., Kehoe, J. L. and Kilic, I. Y. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems*, 33, 2 (2016), 456-481.
[40] Wahyuni, E. D. and Djunaidy, A. *Fake review detection from a product review using modified method of iterative computation framework*. City, 2016.
[41] Heydari, A., Tavakoli, M. and Salim, N. Detection of fake opinions using time series. *Expert Systems with Applications*, 58 (2016), 83-92.
[42] Wang, X., He, K. L. S. and Zhao, J. *Learning to Represent Review with Tensor Decomposition for Spam Detection*. City, 2016.
[43] Hazim, M., Anuar, N. B., Ab Razak, M. F. and Abdullah, N. A. Detecting opinion spams through supervised boosting approach. *PLoS One*, 13, 6 (2018), e0198884.
[44] Budhi, G. S., Chiong, R., Wang, Z. and Dhakal, S. Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. *Electronic Commerce Research and Applications*, 47, 101048 (2021).
[45] Budhi, G. S., Chiong, R. and Wang, Z. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications*, 80 (2021), 13079-13097.
[46] Budhi, G. S., Chiong, R., Pranata, I. and Hu, Z. *Predicting rating polarity through automatic classification of review texts*. City, 2017.
[47] Campbell, C. and Ying, Y. *Learning with Support Vector Machines*. Morgan & Claypool, 2011.
[48] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. *Learning internal representations by error propagation*. MIT Press, City, 1986.
[49] Quinlan, J. R. Induction of decision trees. *Machine Learning*, 1, 1 (March 01 1986), 81-106.
[50] Norvig, P. *How to write a spelling corrector*. City, 2016.
[51] Deng, X., Li, Y., Weng, J. and Zhang, J. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78, 3 (2019), 3797-3816.
[52] Menard, S. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE, Los Angeles, 2010.
[53] Nelder, J. A. and Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135, 3 (1972), 370-384.
[54] Hastie, T. and Tibshirani, R. *Generalized Additive Models*. Chapman and Hall/CRC, United Kingdom, 1990.
[55] Dunteman, G. H. and Ho, M.-H. R. *Generalized Linear Models*. SAGE Publications, Inc., City, 2011.
[56] Dobson, A. J. and Barnett, A. G. *An Introduction to Generalized Linear Models*. CRC Press, Boca Raton, 2008.
[57] Chang, C. C. and Lin, C. J. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 3 (2011), 1-27.
[58] Kingma, D. P. and Ba, J. *Adam: A method for stochastic optimization*. City, 2015.
[59] Hunt, E. B., Marin, J. and Stone, P. J. *Experiments in induction*. Academic Press, New York, 1966.
[60] Rokach, L. and Maimon, O. Z. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing, Singapore, 2007.
[61] Wang, S., Jiang, L. and Li, C. Adapting naïve bayes tree for text classification. *Knowledge & Information Systems*, 44 (2015), 77-89.
[62] Zhang, L., Jiang, L., Li, C. and Kong, G. Two feature weighting approaches for naïve Bayes text classifiers. *Knowledge-Based Systems*, 100 (2016), 137-144.

[63] Jiang, L., Li, C., Wang, S. and Zhang, L. Deep feature weighting for naïve bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, 52 (2016), 26-39.
[64] Manning, C. D., Raghavan, P. and Schuetze, H. *Naïve Bayes text classification*. Cambridge University Press, City, 2008.
[65] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 11 (1998), 2278-2324.
[66] Krizhevsky, A., Sutskever, I. and Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 6 (2017), 84-90.
[67] Yu, Y., Lin, H., Meng, J. and Zhao, Z. Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks. *Algorithms*, 9, 2 (2016).
[68] Simonyan, K. and Zisserman, A. *Very deep convolutional networks for large-scale image recognition*. City, 2015.
[69] Zhu, J., Zou, H., Rosset, S. and Hastie, T. Multi-class AdaBoost. *Statistics and Its Interface*, 2 (2009), 349-360.
[70] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, É. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 85 (May 08 2011), 2825-2830.
[71] Keras *Keras: The Python Deep Learning library*. City, 2019.
[72] Ren, Q., Li, M. and Han, S. Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. *Big Earth Data*, 3, 1 (2019), 8-25.
[73] Xiong, T., Bao, Y., Hu, Z. and Chiong, R. Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. *Information Sciences*, 305 (2015), 77-92.

APPENDIX RESULTS OF THE COMBINATION OF BASE CLASSIFIERS FOR THE MTCE MODEL

Num	MtCE Base Classifier Combination	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	LR-LSVM	87.31	86.70	88.16	87.39	83.33	84.42	90.61	87.16
2	LR-MLP	88.00	86.85	89.44	88.08	84.41	84.23	92.97	88.33
3	LR-DT	86.69	86.96	86.48	86.67	80.82	80.96	91.62	85.91
4	LR-NB	88.56	87.47	90.07	88.72	85.29	86.51	92.21	89.00
5	LSVM-MLP	87.50	85.72	90.02	87.74	86.56	87.05	93.12	89.88
6	LSVM-DT	86.06	86.19	86.01	86.03	82.64	83.84	91.12	87.07
7	LSVM-NB	89.19	88.32	90.43	89.31	84.75	86.85	90.32	88.19
8	MLP-DT	87.81	87.80	87.28	87.50	85.13	83.71	95.45	89.01
9	MLP-NB	89.63	88.52	91.22	89.80	85.31	84.00	95.30	89.12
10	DT-NB	88.50	88.08	89.09	88.53	85.68	85.99	92.73	89.01
11	CNN-LR	85.13	85.54	84.33	84.89	82.59	83.88	90.07	86.75
12	CNN-LSVM	84.88	83.70	86.80	85.08	83.15	85.03	89.20	86.95
13	CNN-MLP	86.19	86.33	86.29	86.22	83.52	83.56	92.74	87.69
14	CNN-DT	81.31	81.96	80.53	81.15	79.59	80.82	89.31	84.75
15	CNN-NB	86.88	86.45	87.60	86.91	85.31	86.04	92.13	88.78

16	LR-LSVM-MLP	87.56	86.22	89.36	87.73	83.13	83.63	92.08	87.46
17	LR-LSVM-DT	86.50	85.61	87.68	86.56	82.61	83.56	91.25	86.89
18	LR-LSVM-NB	89.00	88.68	89.59	89.07	84.58	85.70	91.44	88.34
19	LR-MLP-DT	88.06	87.64	88.66	88.09	81.72	81.82	91.57	86.29
20	LR-MLP-NB	88.50	87.68	89.65	88.59	85.13	85.40	92.81	88.85
21	LSVM-MLP-DT	87.88	87.26	88.90	88.01	84.42	84.66	92.54	88.21
22	LSVM-MLP-NB	89.13	87.86	91.00	89.31	85.13	86.39	91.88	88.79
23	MLP-DT-NB	88.00	87.00	89.30	88.09	84.95	84.25	94.29	88.71
24	CNN-LR-DT	85.25	84.72	86.04	85.30	81.55	81.68	91.92	86.19
25	CNN-LR-LSVM	86.63	86.41	86.62	86.42	80.65	81.80	89.33	85.24
26	CNN-LR-MLP	88.19	87.32	89.33	88.27	84.40	83.75	93.51	88.29
27	CNN-LR-NB	87.50	86.60	88.59	87.52	84.58	85.68	91.33	88.23
28	CNN-LSVM-DT	84.88	85.12	84.72	84.87	83.21	84.92	90.47	87.49
29	CNN-LSVM-MLP	87.75	87.16	88.35	87.71	83.87	84.51	91.67	87.76
30	CNN-LSVM-NB	87.69	87.23	88.02	87.59	85.89	86.61	92.38	89.15
31	CNN-MLP-DT	86.50	86.55	86.25	86.29	83.70	83.51	92.99	87.82
32	CNN-MLP-NB	87.75	86.65	89.17	87.87	84.05	83.73	93.60	88.21
33	CNN-NB-DT	86.81	86.71	87.12	86.81	82.79	82.55	92.70	87.29
34	LR-LSVM-MLP-DT	87.94	87.55	88.16	87.79	82.99	82.63	92.49	87.02
35	LR-LSVM-MLP-NB	89.19	88.77	89.95	89.28	85.66	87.01	91.38	88.94
36	LR-LSVM-DT-NB	88.25	87.83	89.11	88.39	84.06	85.49	90.49	87.75
37	LR-MLP-DT-NB	88.75	88.41	89.25	88.79	84.06	83.85	93.06	88.04
38	LSVM-MLP-DT-NB	88.63	87.70	89.74	88.61	86.38	87.52	92.38	89.67
39	CNN-LR-LSVM-DT	85.94	85.82	86.07	85.87	82.44	83.18	90.71	86.65
40	CNN-LR-LSVM-MLP	86.88	86.69	87.33	86.90	83.17	84.29	90.92	87.29
41	CNN-LR-LSVM-NB	87.25	85.55	89.77	87.52	84.77	86.34	91.00	88.39
42	CNN-LR-MLP-DT	87.38	86.90	88.17	87.49	82.80	82.37	93.17	87.38
43	CNN-LR-MLP-NB	87.75	87.22	88.63	87.84	84.76	84.26	94.15	88.82
44	CNN-LSVM-MLP-DT	87.13	86.76	87.66	87.12	86.21	86.60	92.66	89.48
45	CNN-LSVM-MLP-NB	87.63	87.04	88.48	87.69	85.66	85.59	92.99	89.00
46	CNN-MLP-DT-NB	88.00	87.88	88.21	87.97	84.59	84.50	93.04	88.47
47	LR-LSVM-MLP-DT-NB	88.94	88.38	89.56	88.94	84.24	84.89	92.00	88.12
48	CNN-LR-LSVM-DT-NB	87.88	88.05	87.52	87.69	84.42	85.27	91.48	88.09
49	CNN-LR-LSVM-MLP-DT	87.50	87.10	88.08	87.52	82.07	82.71	91.21	86.61
50	CNN-LR-LSVM-MLP-NB	88.88	88.21	89.83	88.95	86.21	86.45	93.18	89.54
51	CNN-LR-MLP-DT-NB	87.44	86.75	87.98	87.35	85.66	85.57	93.53	89.20
52	CNN-LSVM-MLP-DT-NB	88.25	88.54	87.90	88.16	85.48	85.22	93.55	89.05
53	CNN-LR-LSVM-MLP-DT-NB	87.31	87.02	87.71	87.31	83.17	83.50	92.15	87.40
		Li et al.'s Dataset (Hotel)				Li et al.'s Dataset (Restaurant)			
1	LR-LSVM	85.37	84.23	91.73	87.75	81.13	80.49	82.52	80.59
2	LR-MLP	87.34	85.09	94.30	89.45	85.11	83.99	88.36	85.69
3	LR-DT	84.52	83.82	90.57	87.00	81.34	79.91	83.92	81.53
4	LR-NB	87.39	87.06	91.74	89.30	84.33	82.56	87.35	84.76
5	LSVM-MLP	86.12	84.71	92.67	88.44	85.83	83.98	88.55	86.02

6	LSVM-DT	83.72	83.43	89.47	86.29	81.87	81.45	83.01	81.99
7	LSVM-NB	87.45	87.09	91.75	89.34	85.35	84.51	86.33	85.21
8	MLP-DT	86.01	84.23	92.95	88.31	85.55	86.74	84.42	85.23
9	MLP-NB	87.23	86.69	91.92	89.17	86.04	85.36	87.18	86.15
10	DT-NB	85.74	86.89	88.14	87.50	85.07	85.05	84.76	84.76
11	CNN-LR	83.51	84.05	87.95	85.92	79.84	79.35	80.62	79.56
12	CNN-LSVM	82.98	83.63	87.61	85.56	81.34	83.23	80.11	81.07
13	CNN-MLP	84.68	84.71	89.66	87.03	81.82	81.39	84.25	82.20
14	CNN-DT	81.38	82.87	85.28	84.00	75.63	74.72	77.99	75.96
15	CNN-NB	84.68	86.66	86.73	86.63	83.59	82.63	86.58	84.31
16	LR-LSVM-MLP	86.54	84.99	92.96	88.73	83.33	82.29	84.50	83.27
17	LR-LSVM-DT	85.48	84.70	91.30	87.83	80.83	78.96	83.45	80.75
18	LR-LSVM-NB	85.96	85.20	91.34	88.14	86.58	85.71	88.08	86.41
19	LR-MLP-DT	87.18	85.75	93.40	89.33	84.79	84.37	85.77	84.63
20	LR-MLP-NB	87.18	86.31	92.34	89.19	84.80	83.32	86.30	84.49
21	LSVM-MLP-DT	86.33	84.69	93.17	88.69	82.35	82.10	83.87	82.35
22	LSVM-MLP-NB	86.97	86.09	92.34	89.05	85.32	83.68	88.65	85.75
23	MLP-DT-NB	86.86	86.12	92.02	88.93	84.80	84.00	85.89	84.67
24	CNN-LR-DT	84.52	83.48	91.13	87.10	79.12	77.82	80.86	79.11
25	CNN-LR-LSVM	84.57	83.82	90.67	87.04	79.35	78.24	82.94	79.95
26	CNN-LR-MLP	85.85	84.79	91.96	88.17	83.34	82.42	84.09	82.88
27	CNN-LR-NB	86.54	86.95	90.12	88.47	83.34	82.92	84.77	83.00
28	CNN-LSVM-DT	84.57	84.83	89.24	86.88	79.10	78.69	80.41	79.14
29	CNN-LSVM-MLP	85.11	84.11	91.33	87.50	83.10	82.97	83.82	83.05
30	CNN-LSVM-NB	86.33	86.77	89.95	88.26	78.50	75.48	80.49	77.64
31	CNN-MLP-DT	85.05	84.02	91.18	87.42	82.40	81.95	84.67	82.49
32	CNN-MLP-NB	87.13	86.78	91.49	89.04	84.86	87.11	85.31	86.01
33	CNN-DT-NB	85.16	85.70	88.99	87.29	82.57	84.78	80.46	82.08
34	LR-LSVM-MLP-DT	85.90	84.13	92.97	88.29	83.10	80.86	85.57	82.80
35	LR-LSVM-MLP-NB	86.70	85.12	92.92	88.78	82.32	81.51	84.77	82.80
36	LR-LSVM-DT-NB	86.28	85.79	91.29	88.41	84.35	83.62	86.91	84.73
37	LR-MLP-DT-NB	87.82	86.67	93.26	89.81	84.38	83.61	85.67	84.57
38	LSVM-MLP-DT-NB	87.18	86.13	92.67	89.20	84.57	82.77	87.86	84.89
39	CNN-LR-LSVM-DT	85.11	84.17	91.10	87.44	82.08	81.18	82.93	81.78
40	CNN-LR-LSVM-MLP	85.00	84.51	90.96	87.44	82.10	81.55	84.17	82.46
41	CNN-LR-LSVM-NB	85.85	85.19	91.30	88.09	82.88	82.45	83.32	82.62
42	CNN-LR-MLP-DT	85.43	84.25	91.93	87.87	83.54	82.08	83.21	82.24
43	CNN-LR-MLP-NB	86.54	85.87	91.66	88.62	83.62	82.77	84.95	83.54
44	CNN-LSVM-MLP-DT	85.11	84.34	91.26	87.55	80.37	80.19	82.41	80.66
45	CNN-LSVM-MLP-NB	86.76	85.74	92.15	88.82	83.35	83.36	83.37	83.02
46	CNN-MLP-DT-NB	86.91	86.48	91.43	88.87	84.32	83.74	85.90	84.59
47	LR-LSVM-MLP-NB-DT	87.55	86.19	93.28	89.58	84.63	83.29	87.36	85.06
48	CNN-LR-LSVM-DT-NB	86.28	85.77	91.39	88.44	84.03	82.07	86.59	84.07
49	CNN-LR-LSVM-MLP-DT	85.96	84.83	92.06	88.28	82.13	83.09	81.83	81.64

50	CNN-LR-LSVM-MLP-NB	86.76	85.98	91.98	88.81	83.80	82.59	86.17	83.98
51	CNN-LR-MLP-DT-NB	86.06	85.09	91.85	88.31	84.30	84.98	83.43	84.04
52	CNN-LSVM-MLP-DT-NB	85.74	85.20	91.13	88.01	83.59	82.36	85.31	83.56
53	CNN-LR-LSVM-MLP-DT-NB	85.32	85.38	89.80	87.50	83.09	82.68	83.40	82.87

2. First review: Major revision (7 March 2022)

-----Original Message-----
From: Transactions on Internet Technology <onbehalfof@manuscriptcentral.com>
Sent: Monday, 7 March 2022 8:43 AM
To: Raymond Chiong <raymond.chiong@newcastle.edu.au>
Cc: Reggie.Chentes@spi-global.com
Subject: ACM TOIT Manuscript TOIT-2021-0307 - decision

Dear Raymond Chiong,

We have completed the review process for the paper titled, "A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction." has been completed. The decision is to request the authors undergo a major revision for the manuscript prior before it can be considered further for publication.

Your reviews are listed below. We would suggest that you revise your paper according to the reviewers' comments and resubmit the paper for a second round of reviews. If you wish to revise the paper, your revision due date will be on 17-Apr-2022. Please provide a detailed response to the reviewer comments and indicate (with specificity) which parts of the paper have been modified since the last revision.If you do not intend to submit a revised version of your paper, please let us know so that we can formally close your file. Please be mindful when making your revisions that you still need to maintain the size limitations for papers submitted to TOIT.

Thank you for your submission and we look forward to receiving a future revision. Please feel free to contact the journal administrator Reggie Chentes at Reggie.Chentes@spi-global.com if you have any questions or comments.

Regards,

Prof. Dr. Ling Liu
Editor in Chief, ACM TOIT

Associate Editor's Comments to Author:

Associate Editor
Comments to the Author:
The reviews are to some extend conflict. I will give authors one chance to revise the manuscript and suggest the authors to revise the manuscript carefully and provide detailed revision note.

Reviewer(s)' Comments to Author:

Referee: 1

Recommendation: Reject

Comments:
This paper presents a textual-based ensemble approach to detect fake reviews, i.e., Multi-type Classifier Ensemble (MtCE). MtCE leverages heterogeneous ensemble teams, which are composed of different types of machine learning models, to improve detection performance. Experiments are conducted on four public fake review datasets, which demonstrate that MtCE can effectively detect fraudulent reviews and outperform most of the single base classifiers and other ensemble methods.

Detailed comments are as follows:
Positive Aspects:
1. This paper presents an interesting empirical study and demonstrates that heterogeneous ensembles can improve fake review detection performance.
2. A few critical parameters in ensemble learning have been studied, such as the total number of base classifiers, base classifier type assignments, and voting methods, which is helpful for practitioners to build high-quality ensemble teams for fake review detection.
3. This paper provides detailed background on the problem and the machine learning models used in this study.

Negative Aspects:
1. The proposed MtCE simply leverages heterogeneous ensembles to improve the fake review detection performance. The novelty and contributions of this paper are limited, especially given that (1) ensemble learning techniques have already been applied in fake review detection and (2) it is known that heterogeneous ensembles can improve machine learning or deep learning predictive performance.
2. This paper lacks in-depth analysis and explanation of many interesting observations found in the experiments. For example, this study found different base classifier combinations should be used for different datasets. However, it is unclear how to select the base classifiers for different datasets by using MtCE. Another example is that a weak classifier CNN can significantly improve the overall ensemble performance when it is combined with other stronger classifiers. The reason for such improvements is unclear.
3. The presentation of this paper can be significantly improved. For example, several example reviews can be included to effectively demonstrate the improvements of fake review detection by heterogeneous ensembles. Typos include (1) “yk” in Table 2 and (2) “deep earning” in Line 24 of Page 10.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Heavy

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

Referee: 2

Recommendation: Minor Revision

Comments:
In Section 1 on lines 27-42, a brief overview of ensemble classifiers is provided. However, the related work is not covered well enough. A dedicated section on prior art should be included. In this section, the authors should include more details about the recent studies that use ensembling in natural language processing tasks. Their advantages and drawbacks should also be provided with a thorough comparison with MtCE. In this way, the authors can emphasize the contribution of their work.

In Section 2, having separate subsections to explain the different modules of MtCE would be helpful in terms of organization. Creating classifiers, training process, bootstrapping, testing process... These phases should be described in more detail in different subsections.

In Section 3, Figure 2 is not explained. To improve the flow of the paper, dataset details can be provided after explaining all details of the framework, not before. It can be moved into Section 4.

In Section 4, figures 4-7 have no axis labels. More comments should be included about the results in these figures about the effect of number of classifiers. For example, why are these plots have a fluctuating behavior? What should we infer from these results?

In Section 4, figure 8 has no axis labels. More comments should be included about the results in these figures about the effect of bootstrapping parameter.

In Section 4, figure 9 has no axis labels.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Light

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

Referee: 3

Recommendation: Major Revision

Comments:
Summary
This paper proposes MtCE, an ensemble learning-based method for classifying whether a given consumer review is real or fake. Instead of using a team of homogeneous classifiers (i.e., models generated from the same machine learning algorithm), the proposal uses a combination of classifiers of different algorithms (e.g., logistic regression classifiers plus convolutional neural networks). Evaluated

using four datasets, MtCE outperforms existing approaches. The authors also conducted extensive studies on the effect of hyperparameters of MtCE, giving the readers a sense of how MtCE reacts to different configurations.

Strengths
+ The problem to be addressed by MtCE is well motivated. The authors also explained clearly why MtCE focuses on using textual-based features rather than other available options.
+ The flow charts in the paper are well designed. The readers can easily understand how MtCE and the evaluation process are done.
+ Extensive experiments have been conducted on different hyperparameter settings. The authors clearly explained the strengths and weaknesses of the solution.

Weaknesses
- The technical contributions are unclear and limited. It is recommended to list all contributions at the end of the introduction.
- Related works are missing, even though some relevant papers have been discussed. For example, heterogeneous ensemble methods [1,2] have been proposed and used in practice. They use a team of classifiers produced from different ML algorithms to be the base learners. What is the difference between MtCE and such techniques?
- The authors explained how an ensemble of classifiers can outperform a single classifier but the most important part is missing: how an ensemble of "multi-type" classifiers is better than an ensemble of classifiers of one type.
- While the weaknesses of MtCE and the evaluation method have been clearly identified, they raise concerns about the practicality and the fairness of comparisons.
  -- The performance improvement by MtCE is highly sensitive to the combination of ML algorithms used to generate base learners. This can significantly increase the cost for a proper parameter tuning with, e.g., grid search.
  -- The numbers generated to compare different methods in Table 6 are based on different ways to conduct training/testing/validation.
- More in-depth discussions on the experimental results are recommended. For example, the results in Figure 4 to Figure 7 show an oscillating behavior with an increasing number of base models. This is an interesting result and some intuitive explanation of this behavior will be valuable.
- It is recommended to remove the description of each ML algorithm and the cross-validation process as they can be considered background and a relevant citation can be used instead. More detailed descriptions of MtCE will enhance the writing of this work.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Heavy

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: No

3. Reminder from Transactions on Internet Technology Editorial Office (4 April 2022)

UNIVERSITAS KRISTEN PETRA

Gregorius S. <greg@petra.ac.id>

FW: Reminder: Transactions on Internet Technology - TOIT-2021-0307

Raymond Chiong <raymond.chiong@newcastle.edu.au> Tue, Apr 5, 2022 at 8:50 AM
To: "Gregorius S." <greg@petra.ac.id>

A reminder…

From: Transactions on Internet Technology <onbehalfof@manuscriptcentral.com>
Sent: Monday, 4 April 2022 3:37 PM
To: Raymond Chiong <raymond.chiong@newcastle.edu.au>
Cc: Arriane.Bustillo@straive.com
Subject: Reminder: Transactions on Internet Technology - TOIT-2021-0307

04-Apr-2022

Dear Dr. Chiong:

Recently, you received a decision on Manuscript ID TOIT-2021-0307, entitled "A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction." The manuscript and decision letter are located in your Author Center at https://mc.manuscriptcentral.com/toit.

This e-mail is a gentle reminder that your revision will be due shortly. If it is not possible for you to submit your revision by the deadline, please contact Arriane Bustillo at Arriane.Bustillo@straive.com.

Sincerely,

Transactions on Internet Technology Editorial Office
Arriane.Bustillo@straive.com

4. First Response to the reviewers' and First Revised manuscript files are sent to the Correspondent Author (12 April 2022), and submitted to the Transactions on Internet Technology journal (17 Apr 2022)
    a. First response to the reviewers'
    b. First Revised manuscript with brown font for the revision

UNIVERSITAS KRISTEN PETRA

Gregorius S. <greg@petra.ac.id>

Revision of MtCE paper for ACM ToIT.

1 message

Gregorius Satiabudhi <Gregorius.Satiabudhi@uon.edu.au> Tue, Apr 12, 2022 at 2:30 PM
To: Raymond Chiong <raymond.chiong@newcastle.edu.au>
Cc: "Gregorius S." <greg@petra.ac.id>

Hi Sir,

Please find attached, the revision of MtCE paper for ACM ToIT and the responses file. I also send the Appendix file (no revision here). Thanks.

Kind regards,
Greg

3 attachments

MtCE Appendix.docx
38K

MtCE_Paper_InternetTech (ACM ToIT) rev 090422.docx
1812K

Responses of MtCE paper for ToIT reviews.docx
35K

UNIVERSITAS KRISTEN PETRA

Gregorius S. <greg@petra.ac.id>

MtCE Revised Manuscript

Raymond Chiong <raymond.chiong@newcastle.edu.au> Sun, Apr 17, 2022 at 5:44 PM
To: "Gregorius S." <greg@petra.ac.id>
Cc: Gregorius Satiabudhi <gregorius.satiabudhi@uon.edu.au>

See attached the submitted files

DR RAYMOND CHIONG, SMIEEE
Associate Professor of Data Science & Statistics
School of Information and Physical Sciences
College of Engineering, Science & Environment

O: ES233, Engineering Building S
T: +61 2 4921 7367
E: Raymond.Chiong@newcastle.edu.au
W: http://www.newcastle.edu.au/profile/raymond-chiong

The University of Newcastle (UON)
University Drive, Callaghan, NSW 2308, Australia

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

Top 200 University in the world by QS World University Rankings 2021

CRICOS Provider 00109J

2 attachments

MtCE_Paper_Revised.docx
1795K

Responses of MtCE paper for TolT reviews.docx
26K

Associate Editor Comments to the Author:

The reviews are to some extend conflict. I will give authors one chance to revise the manuscript and suggest the authors to revise the manuscript carefully and provide detailed revision note.

Response:

Thank you for the chance to revise the manuscript. We revised the manuscript as suggested.

Reviewer(s)' Comments to Author:

Referee: 1

Recommendation: Reject

Comments:
This paper presents a textual-based ensemble approach to detect fake reviews, i.e., Multi-type Classifier Ensemble (MtCE). MtCE leverages heterogeneous ensemble teams, which are composed of different types of machine learning models, to improve detection performance. Experiments are conducted on four public fake review datasets, which demonstrate that MtCE can effectively detect fraudulent reviews and outperform most of the single base classifiers and other ensemble methods.

Detailed comments are as follows:

Positive Aspects:

1. This paper presents an interesting empirical study and demonstrates that heterogeneous ensembles can improve fake review detection performance.

2. A few critical parameters in ensemble learning have been studied, such as the total number of base classifiers, base classifier type assignments, and voting methods, which is helpful for practitioners to build high-quality ensemble teams for fake review detection.

3. This paper provides detailed background on the problem and the machine learning models used in this study.

Response:

Thank you for showing us the positive aspects of our paper.

Negative Aspects:

1. The proposed MtCE simply leverages heterogeneous ensembles to improve the fake review detection performance. The novelty and contributions of this paper are limited, especially given that (1) ensemble learning techniques have already been applied in fake review detection and (2) it is known that heterogeneous ensembles can improve machine learning or deep learning predictive performance.

Response:

While ensemble models have already been applied in fake review detection, we proposed our model, the MtCE, to improve detection performance. Also, different to other heterogeneous ensembles, which are usually fixed on the types and total numbers of base classifiers used inside, we proposed a new ensemble where users can customise the types of base classifiers, depending on the problem and their experience and knowledge of the problem. Moreover, while we tested our proposed ensemble model only for fake reviews detection to improve our previous study, we intuitively designed the MtCE to solve general problems of classifications and predictions. We have expressed our intention in section "1. Introduction", at the beginning of paragraph 7, and sub-section "4.3. Datasets" (before 3.1), paragraph 1.

2. This paper lacks in-depth analysis and explanation of many interesting observations found in the experiments. For example, this study found different base classifier combinations should be used for different datasets. However, it is unclear how to select the base classifiers for different datasets by using MtCE.

Response:

Our analysis in paragraph 3, sub-section "5.1. Experimenting with the MtCE for fake review detection" is based only on observation of Table 4. To make it deeper, we add more analysis in the same paragraph from the observation of Table 5 as well.

The best way to do base classifiers candidates selection is to select the candidates from some classifiers that we know perform well on similar topics, then run some experiments using the proposed platform to find the best combination. At the end of paragraph 6 of section 1, sub-section 4.2, paragraph 1, and sub-section 5.1, paragraph 1, we gave examples of choosing the base classifiers candidates (i.e. LR, LSVM, MLP, and CNN) that are performing excellently in our previous studies. However, as we add a comment in section 1, paragraph 6, the base classifiers candidates are better selected based on the problem and the user's experiences of his problem.

Another example is that a weak classifier CNN can significantly improve the overall ensemble performance when it is combined with other stronger classifiers. The reason for such improvements is unclear.

Response:

For CNN discussion (as in the last paragraph of sub-section 5.1), we have never claimed that CNN *"can significantly improve the overall ensemble performance when it is combined with other stronger classifiers"*. We wrote, *"it performed exceptionally well when combined with other classifiers in the MtCE"*. So here, the focus is not that CNN could improve the performance of stronger classifiers, but together with other classifiers in MtCE, the performance of CNN is improved. Moreover, the discussion about CNN in MtCE aims to show that the Deep Learning CNN model could be combined seamlessly with other Machine Learning models and perform well.

3. The presentation of this paper can be significantly improved. For example, several example reviews can

be included to effectively demonstrate the improvements of fake review detection by heterogeneous ensembles. Typos include (1) "yk" in Table 2 and (2) "deep earning" in Line 24 of Page 10.

Response:

Thank you for the revisions. We add a dedicated section, "2. Related Work On Ensemble Models", that discusses the implementation of ensemble models in fake review detection studies, including several heterogeneous ensembles that some researchers in their papers propose. The typos are also fixed.

Additional Questions:

Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Heavy

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

==========

Referee: 2

Recommendation: Minor Revision

Response:

Thank you for the chance to revise our manuscript.

Comments:
In Section 1 on lines 27-42, a brief overview of ensemble classifiers is provided. However, the related work is not covered well enough. A dedicated section on prior art should be included. In this section, the authors should include more details about the recent studies that use ensembling in natural language processing tasks. Their advantages and drawbacks should also be provided with a thorough comparison with MtCE. In this way, the authors can emphasise the contribution of their work.

Response:

Thank you for the revision. A dedicated section, "2. Related Work On Ensemble Models", that discusses recent studies that proposed or implemented ensemble models for fake review detection is added. In this section 2, we also discuss the difference between MtCE to other ensemble models.

In Section 2, having separate subsections to explain the different modules of MtCE would be helpful in

terms of organisation. Creating classifiers, training process, bootstrapping, testing process... These phases should be described in more detail in different subsections.

Response:

Thank you for the suggestion. We split section "3. The MtCE" (before section 2) into several sub-sections. Each sub-section discusses one element/process of MtCE.

In Section 3, Figure 2 is not explained. To improve the flow of the paper, dataset details can be provided after explaining all details of the framework, not before. It can be moved into Section 4.

Response:

Thank you for the idea. Implicitly, we discuss what we do for the framework in section "1. Introduction", paragraph 4. However, to make it more clear, we add a paragraph (paragraph 2) in sub-section "4.1 Framework For Testing" (before 3.2). The datasets sub-section is moved to sub-section 4.3.

In Section 4, figures 4-7 have no axis labels. More comments should be included about the results in these figures about the effect of number of classifiers. For example, why are these plots have a fluctuating behavior? What should we infer from these results?

Response:

We wrote the information about axis labels in figures 4-7 in their caption. However, to make it clearer, we add information in the captions of figures 4-7 with y-labels. Also, we add an intuitive analysis for the fluctuating behaviours of all measurements after 20 classifiers in sub-section "5.2.1 Total Number Of Base Classifiers" (before 4.2.1).

In Section 4, figure 8 has no axis labels. More comments should be included about the results in these figures about the effect of bootstrapping parameter.

Response:

We add the information about the y-axis labels of figure 8 in the caption so that the reader can easily guess the x-axis. We also add analysis for figure 8 in sub-section "5.2.3 Bootstrap Effect" (before 4.2.3).

In Section 4, figure 9 has no axis labels.

Response:

Thank you, we add the information about the y-axis labels in figure 8 in the caption.

Additional Questions:

Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Light

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

==========

Referee: 3

Recommendation: Major Revision

Response:

Thank you for the chance to revise.

Comments:
Summary
This paper proposes MtCE, an ensemble learning-based method for classifying whether a given consumer review is real or fake. Instead of using a team of homogeneous classifiers (i.e., models generated from the same machine learning algorithm), the proposal uses a combination of classifiers of different algorithms (e.g., logistic regression classifiers plus convolutional neural networks). Evaluated using four datasets, MtCE outperforms existing approaches. The authors also conducted extensive studies on the effect of hyperparameters of MtCE, giving the readers a sense of how MtCE reacts to different configurations.

Strengths
+ The problem to be addressed by MtCE is well motivated. The authors also explained clearly why MtCE focuses on using textual-based features rather than other available options.

+ The flow charts in the paper are well designed. The readers can easily understand how MtCE and the evaluation process are done.

+ Extensive experiments have been conducted on different hyperparameter settings. The authors clearly explained the strengths and weaknesses of the solution.

Response:

Thank you for showing the strength of our paper.

Weaknesses
- The technical contributions are unclear and limited. It is recommended to list all contributions at the end of the introduction.

Response:

We add a dedicated paragraph in section "1 Introduction" that discusses all of our technical contributions (see Introduction, paragraph 7).

- Related works are missing, even though some relevant papers have been discussed. For example, heterogeneous ensemble methods [1,2] have been proposed and used in practice. They use a team of classifiers produced from different ML algorithms to be the base learners. What is the difference between MtCE and such techniques?

Response:

We add a new section (section 2) to discuss the related studies in fake review detection that implemented existing ensemble models or proposed new ensemble models (including heterogeneous ensemble models). In the last paragraph of this new section, we show the difference between our proposed ensemble model, the MtCE, and previous models.

- The authors explained how an ensemble of classifiers can outperform a single classifier but the most important part is missing: how an ensemble of "multi-type" classifiers is better than an ensemble of classifiers of one type.

Response:

We have discussed that an ensemble of "multi-type" classifiers is better than an ensemble of classifiers of one type, such as RF, BP, AB and GB in sub-section "5.1. Experimenting With The MtCE For Fake Review Detection" (before 4.1), paragraph 4. To make it more apparent, we add the words "homogenous" in the first sentence and "ensemble" in the second sentence of this paragraph.

- While the weaknesses of MtCE and the evaluation method have been clearly identified, they raise concerns about the practicality and the fairness of comparisons.

-- The performance improvement by MtCE is highly sensitive to the combination of ML algorithms used to generate base learners. This can significantly increase the cost for a proper parameter tuning with, e.g., grid search.

Response:

Thank you for reminding us about this problem. Luckily, in 2020, we proposed PSO based algorithms, namely ML-PSO and PML-PSO [1]. These algorithms are designed to optimise the parameters of ensemble models, such as BP and AB, choose the type of the base classifiers, and optimise the parameters of the candidates of base classifiers. Currently, these algorithms could only handle homogenous ensemble models (i.e. BP and AB), but with a proper modification, we can make it optimising the parameters of MtCE. We mention this plan in the last paragraph of the conclusion section.

-- The numbers generated to compare different methods in Table 6 are based on different ways to conduct training/testing/validation.

Response:

We are aware of these facts and could not do anything to make them all exactly the same since they are results from other publications. Therefore, as in the last paragraph of sub-section "5.3 Comparison To Other Studies On The Same Datasets" (before 4.3), we did not claim that MtCE is better or worse than the results of other studies. We merely present the best results of other studies on similar topics, datasets, and measurements.

- More in-depth discussions on the experimental results are recommended. For example, the results in Figure 4 to Figure 7 show an oscillating behavior with an increasing number of base models. This is an interesting result and some intuitive explanation of this behavior will be valuable.

Response:

Thank you for the suggestion. We add an intuitive analysis for the oscillating behaviours of all measurements after 20 classifiers in sub-section "5.2.1 Total Number Of Base Classifiers" (before sub-section 4.2.1).

- It is recommended to remove the description of each ML algorithm and the cross-validation process as they can be considered background and a relevant citation can be used instead. More detailed descriptions of MtCE will enhance the writing of this work.

Response:

Thank you for the revision. We delete descriptions of each ML algorithm in sub-section 4.2, the CV process in sub-section 4.4 and replace them with relevant citations. The description of MtCE is also detailed in section 3.

Additional Questions:

Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Heavy

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: No

---

[1] BUDHI, G.S., CHIONG, R., and DHAKAL, S., 2020. Multi-level particle swarm optimisation and its parallel version for parameter optimisation of ensemble models: a case of sentiment polarity prediction. *Cluster Computing 23*, 3371-3386. DOI= http://dx.doi.org/https://doi.org/10.1007/s10586-020-03093-3.

A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction

Gregorius Satia Budhi*

School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia, Gregorius.Satiabudhi@uon.edu.au

Informatics Department, Petra Christian University, Surabaya 60236, Indonesia, Greg@petra.ac.id

Raymond Chiong*

School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia, Raymond.Chiong@newcastle.edu.au

Abstract: The financial impact of online reviews has prompted some fraudulent sellers to generate fake consumer reviews for either promoting their products or discrediting competing products. In this study, we propose a novel ensemble model—the Multi-type Classifier Ensemble (MtCE)—combined with a textual-based featuring method, which is relatively independent of the system, to detect fake online consumer reviews. Unlike other ensemble models that utilise only the same type of single classifier, our proposed ensemble utilises several customised machine learning classifiers (including deep learning models) as its base classifiers. The results of our experiments show that the MtCE can adequately detect fake reviews, and that it outperforms other single and ensemble methods in terms of accuracy and other measurements in all the relevant public datasets used in this study. Moreover, if set correctly, the parameters of MtCE, such as base-classifier types, the total number of base classifiers, bootstrap and the method to vote on output (e.g., majority or priority), further improve the performance of the proposed ensemble.

Keywords: Fake review detection, online commerce security, novel ensemble model, machine learning, deep learning.

1 INTRODUCTION

Fake review detection is a hot research topic in natural language processing [55]. A fake review is a consumer review of a product with fictitious opinions written deliberately to sound authentic, for commercial motives; that is, to promote or damage the reputation of the reviewed product [42; 50]. There is a significant possibility that fake reviews distort the actual evaluation of a product [23], create harmful effects and eventually reduce trust in consumer reviews and undermine the effectiveness of the online market [43]. These fraudulent acts occur in response to the vital role of consumer reviews in purchasing behaviour in online markets [68]; consumers trust opinions expressed in online reviews and depend on them to make decisions [4; 14; 39]. Without detection efforts, reviews may be replete with lies, fakes and deceptions, and thus completely useless—or worse [55].

The majority of studies in fake review detection follow three main approaches: those based on the content of the reviews, those based on the behaviour of the reviewers, or a combination of these [3; 29]. Content-based methods focus on the linguistic features of text such as words, parts of speech (POS), n-gram, term frequency (TF) or other linguistic characteristics [21; 27; 55]. The behaviour-based approach focuses more on reviewers' behaviour, such as the user's identity, reviewed product, total number of reviews, ratings given, and duration of reviews [1; 3; 59]. This approach is straightforward, needs only a small number of features, and can deliver good results if properly designed. However, the behaviour-based approach is entirely dependent on additional information, such as metadata, provided by the system. In contrast, the content-based approach is independent of the system, requiring only the review text.

* Corresponding authors' emails:
Raymond.Chiong@newcastle.edu.au, Greg@petra.ac.id, Gregorius.Satiabudhi@uon.edu.au

The content-based approach has two sub-categories. The first creates input features from the review text using Bag of Words (BOW), Word2Vec and skip-gram; thus, the input features for detection are words or terms created from the review text itself, and we call this approach textual-based featuring. The features extracted by this approach are more or less similar, mainly in the form of n-gram terms [14; 21; 27; 42; 63]. This textual-based approach is promising and is used in many studies to extract features from review texts. While independent of the system and needing only the text, this approach usually produces good results [11; 21; 27; 41; 42; 49; 55; 63; 72]. The second form of content-based method attempts to extract information and properties from the text, such as the length of the text, total words and total sentences, in addition to linguistic characteristics of the text, such as POS, subjectivity, complexity, diversity and similarity, and sentiment analysis [12; 26; 28; 57; 65; 66; 71]. This type of content-based featuring is lightweight compared with textual-based featuring; for example, only 80 features in [12] compared with 5,000 features in [11]. This approach, too, is independent of the system since it requires only the text; however, performance is usually relatively poor if not combined with other approaches [12].

In this study, we focus on the textual-based approach. For our experiments, we used Ott et al.'s [49] and Li et al.'s [41] datasets. These datasets are public fake review datasets used in many studies (e.g., see [12; 21; 27; 42; 55; 57; 72]). Before extracting the input features, we implemented text preprocessing such as tokenisation, stopword removal, detection of negation words, correction of elongation words, and POS lemmatisation. To extract the input features, we used the BOW method, n-gram words (from unigram to trigram words) and the count vectorised method. These methods convert a collection of text documents to a matrix of token counts. The token count features are in descending order by TF across the corpus (dataset). To detect fake reviews, we propose a novel machine learning (ML) ensemble model—the Multi-type Classifier Ensemble (MtCE)—that utilises several different types of standard (single) classifiers as its base classifier.

Ensemble classifiers are increasingly used in many different areas of study, such as text analysis [10; 39], computer security [30; 31], environmental science [62], medical research [69], electrical fault detection [47] and spatial data processing [34]. The main idea behind ensemble classifier models is to combine multiple single classifiers to produce a combination that outperforms any single classifier itself [8; 56]. In general, ensemble classifier models can be classified into four groups: bagging, boosting, stacked generalisation, and the mixture of experts [52]. Bagging, also called bootstrap aggregating, accumulates multiple predictors/classifiers and runs a plurality vote when predicting classes. These base classifiers are differentiated using a bootstrap technique to replicate the learning set [5]. The Random Forest (RF) and Pasting Small Votes ensemble models are a variation of bagging [6; 7]. The second group of ensemble models is boosting. Similar to bagging, boosting is also an aggregation of multiple classifiers. However, instead of using bootstrap, it uses the boosting technique to train the base classifiers. This technique can convert weak learning algorithms to achieve arbitrarily high accuracy [60]. Adaptive Boosting (AB) [24] and Gradient-Boosted Trees (GB) [25] are two well-known algorithms that use this technique. In stacked generalisation [67], the classifiers are stacked to one another so that the output of some first-level base classifiers becomes the input of a second-level meta classifier. The combination of expert methods [32] combines the weighted output of several base classifiers in the consecutive combiner.

While promising, we see the potential for improvement from these ensemble classifiers; they typically use only the same type of single classifier as the base classifier. To differentiate the output of each base classifier, they implement bootstrapping or boosting of the training sample inputs via bagging or boosting, stacking the classifiers, or applying different weights for their output. In contrast, our proposed MtCE combines several types of single classifiers working together as an ensemble model. As every single classifier has strengths and weaknesses in classification or detection, by combining different types of single classifiers in an ensemble, we use the strength of one classifier to 'cover' for the weakness of the other classifiers, and vice versa. The types of base classifiers ensembled in our MtCE can be customised by the user based on the problem and their experience and knowledge of this problem. To further increase the performance of our model, we implemented the bootstrap technique [5] and the method to vote on output (majority or priority) to produce the final output. For the base classifiers of our proposed ensemble model, we implemented several standard single classifiers that performed well in previous research on text mining [9-12; 16; 17], including the Logistic Regression (LR) [46], Linear-kernel Support Vector Machine (LSVM) [13; 15], Multi-Layer Perceptron (MLP)

[58], and Convolutional Neural Network (CNN) [40]. In addition, two other classifiers that are usually applied as the base classifiers in several ensemble models, namely Decision Tree (DT) [53] and Naïve Bayes (NB) [44], were also used as base classifiers. The MtCE was compared with ensemble models widely used in the literature, including Bagging Predictors (BP) [5], RF, AB and GB.

Theoretically, our work contributes to the artificial intelligence research domain, especially ML, by proposing a novel ensemble approach that can solve classification, detection and prediction problems. Most ensemble models in the literature either (1) have only one type of base classifier (such as RF, AB, BP and GB) or (2) have a small combination of fixed types and number of base classifiers that are designed to solve a particular problem [33; 55; 63; 64]. In contrast, our proposed model (the MtCE) provides the freedom to select the types and number of base classifiers depending on one's requirements. Bootstrapping is included in the MtCE process to improve the model's stability and accuracy. Finally, a priority threshold approach is introduced, in addition to majority voting, to decide the final result; this priority threshold approach can improve the recall and overcome the imbalanced issue.

From a practical perspective, this work also contributes to addressing online commerce security problems—we have successfully implemented a model that can detect fake online consumer reviews with better results than previous research. Product reviews from buyers are commonly used as a guide by most consumers to make purchasing decisions on electronic commerce these days [9]. Reliable and effective detection of fake reviews will increase the trustworthiness of online commerce[10]. On the other hand, sellers use consumer reviews to evaluate the brand perception and consumers' satisfaction [22]. For this reason, maintaining the quality of product reviews is vital. Therefore, fake review detection is an urgent task and a high priority for online commerce portal providers.

The rest of this paper is organised as follows. In the next section, we explain the design of our proposed model, the MtCE. After that, we discuss how we conducted experiments to test the performance of the MtCE, such as the datasets we used, testing framework, details of the textual-based approach, types of base classifiers, and the measurements. In Section 4, we discuss the results and analyses, and finally, we draw conclusions and outline the possibilities for future work.

2 RELATED WORK ON ENSEMBLE MODELS

The previous section provided a concise introduction to both fake review detection and ensemble models. This section discusses recent fake review detection studies from the literature (2016 to the present) that have proposed new ensemble models or implemented existing ensemble models for detection of fake reviews (see Table 1).

The objective of most previous studies has been to propose feature extraction approaches for the input of standard ML and deep learning (DL), including their ensemble models. As discussed in Section 1, these feature extraction approaches can either be textual-based, which creates features from the words/terms of the review text itself [11; 20; 21; 72]; or content-based, which creates features from the text and extract features from the text's information, properties, sentiment polarities, and characteristics [2]; or behaviour-based, which emphasises the behaviour of the reviewers rather than their reviews [3; 19; 39; 45].

The above approaches have also been combined to improve the detection results [12; 26; 38; 61; 71]. These studies investigated existing ML, DL and ensemble models (i.e. AB, BP, RF, and GB) to detect fake reviews using features proposed using the combined approaches. The base classifiers of these ensemble models are one type; for example, the RF utilises decision trees for its base classifiers, and GB utilises regression trees. While the type of the base classifier in such ensemble methods can be changed, the replacement must still be one type, i.e., it is impossible to mix more than one type of base classifier for AB and BP.

Some studies in the literature have also proposed novel ensemble models. For example, Sun et al. [63] proposed a bagging style of 2 SVMs and a CNN to detect fake reviews. They utilised special SVMs, namely $BIGRAMS_{SVM}$ and $TRIGRAMS_{SVM,}$ to detect terms features from the review text input, with a unique CNN model (Product Word Composition Classifier (PWCC) to capture product-related review features. The output of these three classifiers were processed in a bagging style method to produce the final detection result. Similarly, a forest of decision trees, called

the Neural Autoencoder Decision Forest, was proposed by Dong et al. [19] to detect fake reviews. Their ensemble was designed by combining the Deep Neural Decision Tree (proposed by Kontschieder et al. [37]) with the Autoencoder method. An integration of several CNNs and Gated Recurrent Neural Networks (GRNNs) were introduced by Ren and Ji [55] in 2017. The CNNs were implemented to produce continuous sentence vectors from words, then handled by several forward and backward GRNNs to produce document representations of the reviews. A unique model of CNN, namely bfGAN, was proposed by Tang et al. [64] in 2020, and combined with SVM for fake review detection. The authors claim that their model can effectively detect cold-start spam/fake reviews; the cold start problem is when a new reviewer posts a review. Javed et al. [33] proposed an ensemble of three classifiers (CNNs) for fake review detection. Each classifier detects different feature groups: a textual-based, a content-based (non-textual), and a behaviour-based group of features; a voting mechanism is the used to produce the final detection. The above-discussed approaches are similar in terms of one important aspect: they have proposed a specific ensemble model with specific types and total number of base classifiers.

Our proposed ensemble model, the MtCE, is different. In the MtCE, diverse base classifiers can be mixed together so that the strength of one type of base classifier can overcome the weaknesses of other types of base classifier and vice versa. The total number of base classifiers and the number of each type of base classifier can also be configured freely based on specific requirements. Our model also applies the bootstrapping technique to ensure stability and improve accuracy. For the final classification result, we introduce a priority threshold technique that prioritises a particular class in conjunction with common majority voting that is usually applied by other ensembles. This priority threshold technique can overcome the imbalanced problem when applied to the scarce/rare class as well as improve the recall in binary classification if applied to the main (positive) class.

Table 1. An overview of related work (those algorithms in italic are ensemble models)

Author	**Year**	**Featuring type**	**Algorithm**
Sun et al. [63]	2016	Textual-based	*Bagging of 2 SVMs and PWCC*
Zhang et al. [71]	2016	Content- and behaviour-based	NB, SVM, DT, *RF*
Etaiwi and Naymat [21]	2017	Textual-based	NB, SVM, DT, *RF, GB*
Ren and Ji [55]	2017	Textual-based	SVM, GRNN, Recurrent Neural Network (RNN), CNN, *CNN-GRNN (Integrated)*
Dong et al. [19]	2018	Behaviour-based	*Neural Autoencoder Decision Forest*
Hazim et al.[26]	2018	Content and behaviour-based	*Generalised Boosted Regression Model (GBM) Gaussian, GBM Poisson, GBM Bernoulli, GB, AB*
Kumar et al. [39]	2018	Behaviour-based	LR, k-NN, NB, SVM, *RF, AB*
Zhang et al. [72]	2018	Textual-based	Recurrent CNN, SVM, CNN, *CNN-GRNN*
Barbado et al. [3]	2019	Behaviour-based	LR, DT, Gaussian NB (GNB), *RF, AB*
Martens & Maalej[45]	2019	Behaviour-based	DT, MLP, LSVM, RBF kernel SVM (RSVM), GNB, *RF*
Tang et al. [64]	2020	Behaviour-based	CNN, *CNN-bfGAN + SVM*
Alsubari et al. [2]	2020	Content-based	DT, *RF, AB*
Budhi et al. [11]	2021	Textual-based	LR, MLP, LSVM, *RF, AB, BP*
Budhi et al. [12]	2021	Content- and behaviour-based	DT, LR, LSVM, MLP, CNN, *RF, GB, AB, BP*
Elmogy et al. [20]	2021	Textual-based	LR, NB, k-NN, SVM, *RF*
Javed et al. [33]	2021	Textual-, content-, and behaviour-based	*Ensemble of 3 CNNs*
Shan et al. [61]	2021	Content- and behaviour-based	DT, SVM, NB, MLP, *RF*
Kumar et al. [38]	2022	Textual-, content-, and behaviour-based	Long short-term memory, Artificial Neural Network, RNN, RSVM, LR, k-NN, NB, *RF, GB, Light GB Method*

3 THE MTCE

The MtCE is a novel ensemble of classifiers developed in light of the fact that every single classifier has its strengths and weaknesses (see Figure 1). This ensemble is proposed based on the following idea:

> *Every single classifier has been created with a different method, formulation and purpose, and thus, has strengths and weaknesses in different spots; for example, (1) for the same dataset, each classifier may correctly classify the different set of records; (2) a classifier is usually created to solve one or two problems, and not tested for the case of other problems; and (3) some classifiers perform well for smaller datasets but not for bigger ones, or vice versa. Therefore, by combining different single classifiers in one ensemble, the strengths of one classifier may 'cover' for the weaknesses of another.*

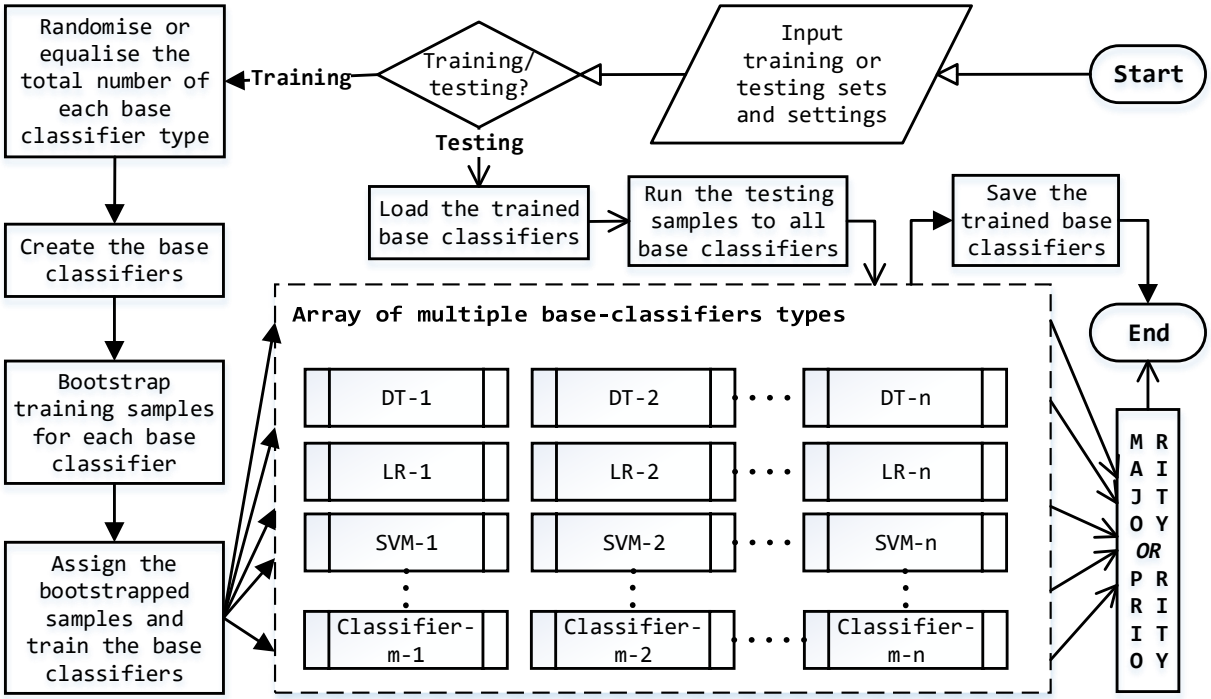

Figure 1: Design of the MtCE

In Figure 1, we depict two processes: the training process (the full head arrows ➔) and the testing process (the line head arrows →). All processes begin with inputting the training or testing sets and the parameter settings of the MtCE (see hollow head arrows -▷).

3.1 BASE CLASSIFIERS OF THE MTCE

Base classifiers here include several classifier objects that are utilised within an ensemble model. These classifiers will be run jointly in a specific way (depending on the ensemble design) to produce results. These results are then processed to produce the final result of the ensemble model. For our MtCE, first the user will decide on the types of base classifiers and their total numbers. After that, for the training process, the user determines if the base classifiers will be created in equally split numbers between the types setting or in random fashion. For example, suppose the total base classifiers are 10 of 3 types (LR, DT, NB); an equally split setting will create 4 LR, 3 DT and 3 NB, while for a random setting, each classifier type will be randomised from 1 to 8 (total classifiers – (total types – 1)) classifiers. While splitting the number of base classifier types equally is more sensible if the user does not know the exact nature of the problem at hand, randomising them sometimes could give better results. Moreover, with a simple modification in the implementation phase, the user can also set the exact different total number of each base classifier type, supposed

they have a reason to do so. In this study, we did not test the effect of setting exact numbers of each type of base classifier since, for our problem, it will become similar to the randomised setting.

3.2 BOOTSTRAP SAMPLING

After creating the base classifiers, each base classifier is assigned a subset of training samples based on the bootstrap sampling process. Bootstrapping the samples could improve the stability and accuracy of the model [5]. The bootstrap sampling method draws sample data repeatedly with replacement from the source (data), based on the percentage setting [5; 35]. After the training process for each base classifier finishes, the weights and parameters of all base classifiers are saved in a file.

3.3 DETERMINING THE FINAL OUTPUT

The testing process is straightforward. After inputting the testing samples and loading the previously trained base classifiers, all samples are run with the base classifiers. The final output will be determined by one of two processes: the majority vote or the priority class threshold. The majority vote chooses the majority class outputs from the results of the base classifiers. If the majority votes of all classes are equal, the final result is the smallest class in the corpus. The priority class threshold provides a final result based on the priority class chosen in the setting; that is, if the positive class is chosen as the priority, then, if the positive class outputs from base classifiers are the same as or more than the threshold, the final output is a positive class. The priority class threshold can be set from absolute, which needs only one base classifier to pick the chosen class, to the maximum threshold before it becomes a majority vote (i.e. more than 50% + 1 in a binary classification problem). This priority class threshold mechanism will focus on the class that is prioritised and will increase the performance of detection of this class.

4 TESTING THE PERFORMANCE OF THE MTCE

4.1 FRAMEWORK FOR TESTING

To evaluate our proposed ensemble, the MtCE, we implemented the comparison framework that we used previously to investigate classifiers for sentiment polarity detection of consumer reviews [10]. In this study, we modified the framework so that it could be applied to test the MtCE for fake review detection (see Figure 2). We used this framework to test different settings of the MtCE using similar datasets and comparing their performances. We ran the same test using several single classifiers as base classifiers of the MtCE and several well-known ensembles for comparison.
As can be seen in Figure 2, after a review text is loaded to a string, it is processed in the preprocessing subroutine to remove unnecessary words and corrections. Features are then extracted from the texts using a textual-based approach, as discussed in Section 1. After combining with the designated targets, these feature-target vectors are split into 10 folds for the cross-validation process. This same 10-fold setting is then trained and tested on several different combinations of MtCE models. Afterwards, the measurement results of these different MtCE models are compared to determine the best setting of the MtCE for fake review detection.

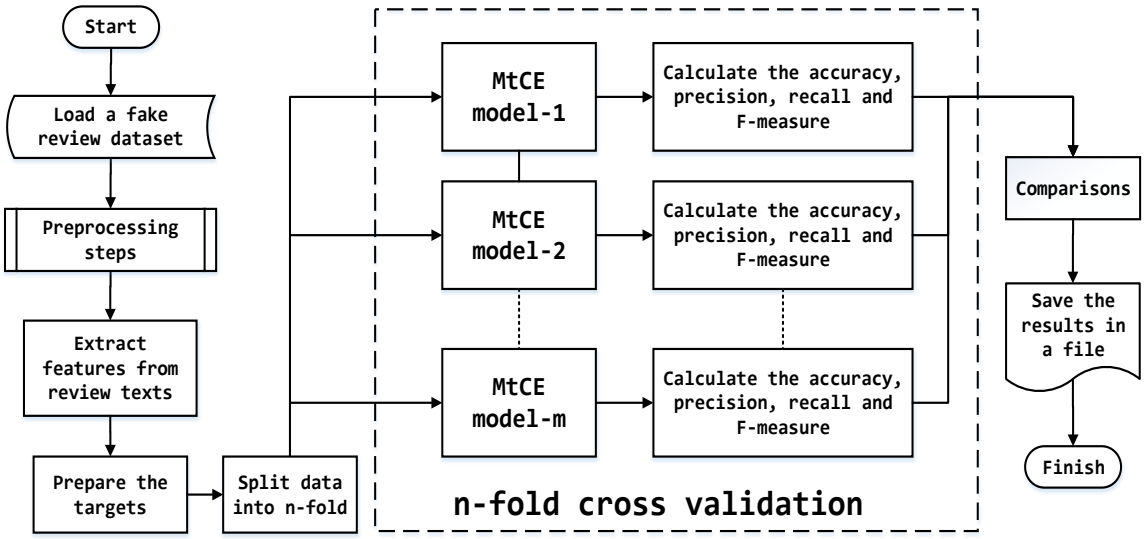

Figure 2: Framework to test and compare the MtCE

In the preprocessing steps (see Figure 3), we implemented several methods as follows:

a. removal of punctuation, numbers and stopwords
b. correction of spelling errors, elongation words and negative words
c. POS tagging and lemmatisation.

Punctuation and number removal is a necessary step for a textual-based approach since the input features of this approach are the words. Therefore, we needed to remove the non-word parts of the text reviews. Stopword is a term for words commonly used in English sentences such as 'the', 'is', 'at', 'which' and 'on'. These words can be removed from the text before it is used as a training record. These kinds of words may mislead detectors in recognising the trait words of a class because they are often the most common words in a corpus.

Misspelled words create unnecessary issues for detection since the detector will recognise them as different words from their correct forms. Like misspelled words, elongation words such as 'yesss', 'fiiine' and 'yoouu' can also increase the diversity of words in the training example and make the classifiers harder to train. We utilised Peter Norvig's code for spelling correction [48] to correct misspelled and elongation words. Negative words have many forms, depending on the grammar, but their purpose is similar—to change a positive sentence to be negative. Therefore, we shifted all negative words to their primary form to reduce the diversity of words.

POS tagging categorises words by their syntactic function, such as noun, pronoun, adjective, verb and adverb. This step is essential because it puts the words in their context [21] so that in the lemmatisation process, we can lemmatise the word to the correct context. Lemmatisation is a method for changing the word back to its basic form; POS lemmatisation returns the chosen word to its basic syntactic function. This step reduces the diversity of words inside the dataset and makes it easier to recognise.

Because we extracted the input features directly from social media messages, we used the BOW feature extraction method commonly used for textual-based features [18]. This method works by decomposing the entire text into a group of singular words. Similar to previous work [11], in this study, we used it in combination with the n-gram technique to capture singular words and terms that convey one meaning but are composed of multiple words. For terms, the BOW checks for the existence of a contiguous sequence of n words from the given text sample. This study limited this checking to trigrams since sequences of more than three words are rare in real-world texts. After decomposing the text, the terms were sorted based on their frequency across the corpus.

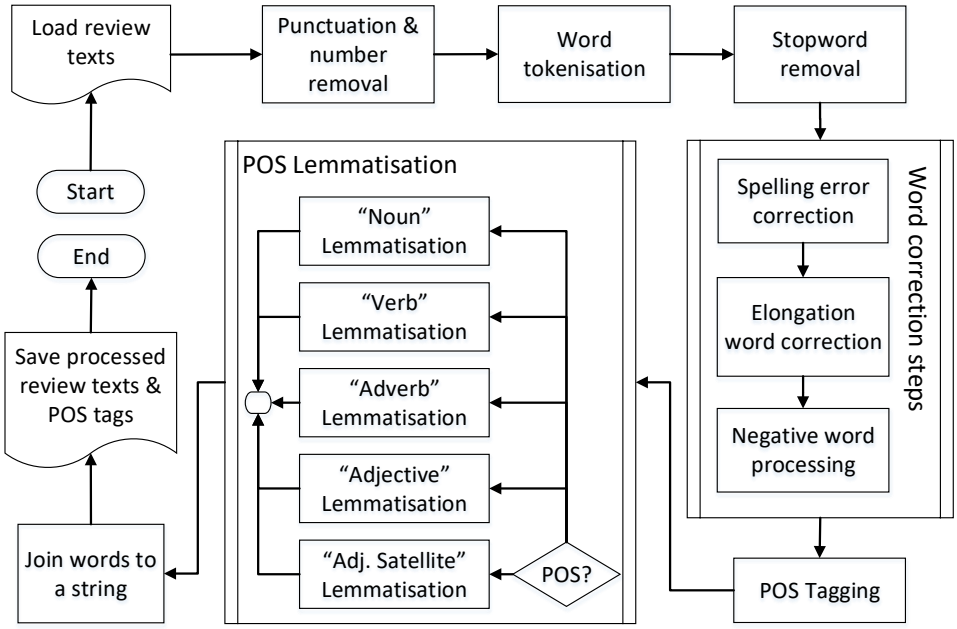

Figure 3: Preprocessing steps [11]

4.2 BASE CLASSIFIERS AND ENSEMBLE CLASSIFIERS FOR EXPERIMENTAL PURPOSE

As mentioned in Section 2, our proposed ensemble can apply multiple types of single classifiers. While in theory, any single classifier could be used as the base classifier of the MtCE, to test its performance, we investigated several combinations of ML and DL that performed well in previous research on text mining [9-12; 16; 17], which include the LR, LSVM, MLP, and CNN models. We also investigated two other classifiers, DT and NB, since these two models are commonly used as default base classifiers in some ensemble models, such as the RF [7], BP [5] and AB [24]. For comparison purposes, we tested some ensemble models BP, RF, AB and GB [25].

We built all ML classifiers and ensembles, except the CNN, using scikit-learn components [51]; the CNN was built with Keras [36] wrapped with scikit-learn components (scikit-learn does not provide a CNN component but a way to wrap DL components of Keras to be used with or within scikit-learn). To ensure the results could only be affected by our model implementation and not by modifying classifier parameters, we used the default parameters for all single classifiers used as base classifiers and the ensemble models.

4.3 DATASETS

Intuitively, our proposed ensemble classifier can be applied to a wide range of problems, similar to other ML ensemble classifiers. However, to test the performance of the MtCE, we conducted experiments using four public fake review datasets (see Table 2 for additional details). The public datasets from Ott et al. [33] and Li et al. [35] were created using domain experts, such as Amazon's Mechanical Turk service (Turkers) and hotel employees, to provide false/fake reviews for some online services. Li et al.'s hotel dataset is an extension of Ott et al.'s dataset.

Table 2: Statistics for several public datasets

Dataset Author	Domain	Total Records	Fake		Genuine	
			Total	%	Total	%
Ott et al. [49]	Hotel (Various)	1600	800	50.00	800	50.00
Li et al. [41]	Doctor (Various)	558	357	63.98	201	36.02
	Hotel (Various)	1880	1080	57.45	800	42.55
	Restaurant (Various)	402	202	50.25	200	49.75

4.4 CROSS-VALIDATION AND MEASUREMENTS

We evaluated the performance of the MtCE using n-fold Cross-Validation (CV), which is widely applied to evaluate the generalisation performance of an algorithm [30], to reduce bias between the dataset and the training or testing set [54], and avoid overfitting [70]. We implemented scikit-learn [51] for performance measurements of our experiments. These measurements and their respective formulas can be found in Table 3.

Table 3: Measurement functions and formulas

No	Name	Sklearn Function	Equation
1	Accuracy	accuracy_score()	$A(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{k=0}^{n_{samples}-1} 1(\widehat{y_k} = y_k)$ where y is the set of predicted pairs, $\hat{y}$ is the set of true pairs and $n_{samples}$ is total samples.
2	Precision	precision_score()	$P(y_k, \widehat{y_k}) = \frac{tp}{tp + fp}$ where k is the set of classes, y_k is the subset of y with class k, tp is true positive, and fp is false positive.
3	Recall	recall_score()	$R(y_k, \widehat{y_k}) = \frac{tp}{tp + fn}$ where fn is false negative.
4	F-measure/F1	f1_score()	$F_1(y_k, \widehat{y_k}) = 2 * \frac{P(y_k, \widehat{y_k}) * R(y_k, \widehat{y_k})}{P(y_k, \widehat{y_k}) + R(y_k, \widehat{y_k})}$

5 RESULTS AND DISCUSSION

5.1 EXPERIMENTING WITH THE MTCE FOR FAKE REVIEW DETECTION

The first group of experiments aimed to test the performance of the MtCE for fake review detection. For the base classifiers, we implemented CNN, LR, LSVM and MLP, which were performing well in our previous studies [9-12]. We also implemented DT and NB, which were used as default base classifiers in many ensembles, such as BP, AB and RF. These base classifier parameters are the defaults of the scikit-learn components used to implement these classifiers [51]. As shown in the Appendix, we tested all possibilities from 2-, 3-, 4- to 5-combination and presented 11 selected combinations in Table 4. Besides combining base classifiers, the other settings were fixed: total base classifiers = 15, shared equally for each type; bootstrap = 0.5; and final output = majority vote. The datasets we used in the experiments were Ott et al.'s dataset [49] and all three of Li et al.'s datasets [41]. All experiments were conducted using the 10-fold CV method. For the comparison, we also predicted the datasets using single classifiers as above and several well-known ensemble classifiers (RF, BP, AB and GB). The results of the single and ensemble classifiers are provided in Table 5.

Table 4: Results of the combination of base classifiers for the MtCE model (selected)

Num[1]	MtCE Base Classifier Combination	Ott et al.'s Dataset[2]				Li et al.'s Dataset (Doctor)[2]			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
2	LR-MLP	**88.00**	*86.85*	**89.44**	**88.08**	84.41	*84.23*	**92.97**	88.33
5	LSVM-MLP	87.50	85.72	**90.02**	**87.74**	**86.56**	**87.05**	**93.12**	**89.88**

7	LSVM-NB	**89.19**	**88.32**	**90.43**	**89.31**	*84.75*	86.85	**90.32**	88.19
9	MLP-NB	**89.63**	**88.52**	**91.22**	**89.80**	*85.31*	*84.00*	**95.30**	*89.12*
18	LR-LSVM-NB	**89.00**	**88.68**	**89.59**	**89.07**	84.58	85.70	**91.44**	88.34
22	LSVM-MLP-NB	**89.13**	87.86	**91.00**	**89.31**	85.13	86.39	91.88	88.79
26	CNN-LR-MLP	**88.19**	**87.32**	**89.33**	**88.27**	84.40	83.75	**93.51**	88.29
35	LR-LSVM-MLP-NB	**89.19**	**88.77**	**89.95**	**89.28**	85.66	87.01	91.38	88.94
37	LR-MLP-DT-NB	**88.75**	**88.41**	89.25	**88.79**	84.06	83.85	**93.06**	88.04
47	LR-LSVM-MLP-DT-NB	**88.94**	**88.38**	**89.56**	**88.94**	84.24	84.89	92.00	88.12
50	CNN-LR-LSVM-MLP-NB	**88.88**	**88.21**	**89.83**	**88.95**	86.21	86.45	**93.18**	89.54
		Li et al.'s Dataset (Hotel)[2]				Li et al.'s Dataset (Restaurant)[2]			
2	LR-MLP	**87.34**	85.09	**94.30**	**89.45**	85.11	83.99	**88.36**	85.69
5	LSVM-MLP	86.12	84.71	**92.67**	**88.44**	**85.83**	83.98	**88.55**	**86.02**
7	LSVM-NB	**87.45**	87.09	**91.75**	**89.34**	85.35	84.51	86.33	85.21
9	MLP-NB	87.23	86.69	**91.92**	**89.17**	86.04	85.36	87.18	**86.15**
18	LR-LSVM-NB	85.96	85.20	**91.34**	88.14	**86.58**	85.71	**88.08**	**86.41**
22	LSVM-MLP-NB	86.97	86.09	**92.34**	**89.05**	85.32	83.68	**88.65**	85.75
26	CNN-LR-MLP	85.85	84.79	**91.96**	88.17	83.34	82.42	84.09	82.88
35	LR-LSVM-MLP-NB	86.70	85.12	**92.92**	88.78	82.32	81.51	84.77	82.80
37	LR-MLP-DT-NB	**87.82**	86.67	**93.26**	**89.81**	84.38	83.61	85.67	84.57
47	LR-LSVM-MLP-DT-NB	**87.55**	86.19	**93.28**	**89.58**	84.63	83.29	87.36	85.06
50	CNN-LR-LSVM-MLP-NB	86.76	85.98	**91.98**	**88.81**	83.80	82.59	86.17	83.98

[1] The number here is associated with the number in Appendix
[2] **Bold-green** = the MtCE result is higher than that of all base classifiers in Table 5; *Italic-red* = the MtCE result is higher than all base classifiers; Normal-black = at least one base classifier result is higher than that of the MtCE.

Table 5: Results of single and ensemble classifiers

Num.	Classifiers	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	CNN	79.63	81.48	77.36	79.13	67.76	70.64	88.40	77.02
2	LR	87.13	87.05	87.25	87.11	83.35	85.48	89.71	87.26
3	LSVM	85.88	85.67	86.21	85.89	83.13	86.28	87.80	86.91
4	MLP	87.56	87.05	88.37	87.66	85.67	85.89	92.94	89.16
5	DT	71.50	72.21	69.71	70.83	65.79	72.98	74.01	73.32
6	NB	88.50	87.97	89.45	88.63	87.12	90.68	88.98	89.68
7	RF	87.00	87.44	86.62	86.92	76.35	74.05	97.24	83.98
8	BP	75.00	74.23	76.79	75.41	74.19	74.50	90.21	81.45
9	AB	80.06	79.64	80.82	80.09	74.93	79.56	81.28	80.16
10	GB	82.81	83.67	81.63	82.56	76.52	77.08	90.63	82.99
		Li et al.'s Dataset (Hotel)				Li et al.'s Dataset (Restaurant)			
1	CNN	78.83	82.48	80.63	81.29	71.63	71.66	73.20	71.75
2	LR	85.59	85.50	90.25	87.77	81.59	81.03	84.52	81.74
3	LSVM	84.20	84.59	88.26	86.33	82.11	79.93	84.40	81.53
4	MLP	86.12	85.89	90.82	88.24	85.56	84.85	87.69	85.73

5	DT	68.88	73.81	71.15	72.38	60.24	60.34	63.27	61.11
6	NB	87.29	89.73	87.92	88.78	86.08	86.36	86.63	86.10
7	RF	81.97	80.78	90.28	85.17	81.34	78.31	87.86	81.88
8	BP	75.96	76.35	84.14	80.03	71.65	71.08	74.51	71.95
9	AB	79.89	80.77	84.87	82.67	70.89	70.64	70.35	70.38
10	GB	80.48	78.83	89.94	84.00	76.12	77.43	75.46	75.78

Comparing Table 4 with Table 5, some combinations of base classifiers in the MtCE exceed the performance of all of the base classifiers (see the bold-green scores in Table 4). For example, accuracy of the MtCE(LR-MLP) for Ott et al.'s = 88%; in comparison, in Table 5, accuracy of LR and MLP are 87.13% and 87.56%. However, not all combinations improved and supported each other as we hoped. A few weak classifiers degraded the performance of stronger classifiers when combined with these. This gave rise to the result that the performance of the MtCE was in the middle of the performance of its base classifiers; for example, LSVM accuracy for Ott et al.'s = 85.88% while MLP = 87.56%, and the combination of both in MtCE(LSVM-MLP) = 87.50%. This implies that the LSVM degraded the performance of MLP, and given this, rather than implementing our proposed ensemble, it would be better to use the MLP directly. However, in the same case, MtCE(LSVM-MLP) for Ott et al.'s improved the recall, from 86.21% (LSVM) and 88.37% (MLP) to 90.02%. In a case such as this, using our ensemble would be acceptable since the recall in binary classification/detection is the same as the accuracy of the positive class or the main class (in our problem, the fake review class). Thus, high recall means the ability of the ensemble to detect positive classes is also high.

Another finding from this set of experiments is that different datasets might need different base classifier combinations. For example, the MtCE(MLP-NB) that performed best for Ott et al.'s dataset performed worst for Li et al.'s doctor dataset (increasing the recall but decreasing all other measurements). The best combinations for Li et al.'s doctor dataset was MtCE(LSVM-MLP); for Li et al.'s hotel dataset was MtCE(LR-MLP-DT-NB); and for Li et al.'s restaurant dataset was MtCE(LR-LSVM-NB). A closer inspection of Table 5 shows that the base classifiers of MtCE's best combination for a particular dataset mostly performed well on the same dataset too, when they were used in stand-alone modes. For example, on Ott's MtCE(MLP-NB), the MLP and NB also performed well on Ott's dataset, although they were not as good as when combined in the MTCE. This observation is valid on other combinations for other datasets, except the DT on Li et al.'s hotel dataset, which was not performing well by itself but performed best in combination with other base classifiers in the MtCE. This indicates that while combining good classifiers could produce better results, infusing a weak classifier sometimes could also boost the results of MtCE.

We found that the MtCE performed better than other ensembles of homogenous classifiers, as indicated by comparing the results in Table 5 (numbers 7–10) to the MtCE results in Table 4. These facts suggest that, when performed correctly, the combination of multi-type classifiers could cover the weakness of each and outperform the combination of a homogenous type of ensemble classifiers such as RF, BP, AB or GB. However, for the cases of RF, BP and AB, we suspect the cause is also because these ensembles use DT as their base classifiers/predictors. DT's performance was not good on all datasets, which therefore affected the performance of its ensemble variants.

While deep learning CNN as a stand-alone classifier did not perform as well as other single classifiers, somewhat surprisingly, it performed exceptionally well when combined with other classifiers in the MtCE. This increase in accuracy was significant; for example, MtCE(CNN-LR-LSVM-MLP-NB) is 7.93% in Li et al.'s hotel dataset to 18.45% in Li et al.'s doctor dataset. The improvement in recall was also good, with a minimum of 4.78% in Li et al.'s doctor dataset to 12.97% in Li et al.'s restaurant dataset. Therefore, the deep learning CNN could be combined perfectly with other ML single classifiers as base classifiers for the MtCE; it gave good results and improved all other base classifiers' performances as well.

5.2 EFFECTS OF MTCE PARAMETERS

In this section, we discuss several input parameters of the MtCE that may affect the performance of this proposed model. All of the settings are the same as in Section 5.1 except the parameter we are testing. For these experiments, we chose four base classifier combinations of the MtCE in Table 4—that is, MtCE(LSVM-MLP), MtCE(LSVM-NB), MtCE(LR-LSVM-NB), MtCE(LR-MLP-DT-NB) and MtCE(LR-LSVM-MLP-DT-NB)—as these are the five combinations of base classifier types that have performed the best across all four datasets. We ran all experiments for the parameter testing using Ott et al.'s dataset only, since these five MtCE combination settings perform well in this dataset (see the bold-green scores in Table 4). However, after finding the ideal set of parameters, we ran them with the other datasets and compared the results with our previous approach implemented on the same datasets.

5.2.1 TOTAL NUMBER OF BASE CLASSIFIERS

First, we investigated the effect of base classifiers' total numbers. Here, we ran experiments on all five combinations with the total number of base classifiers varying from 5 to 100. Results can be seen in Figure 4 for accuracy, Figure 5 for precision, Figure 6 for recall and Figure 7 for F-measure. From these figures, we can see that accuracy, precision, recall and F-measure were increasing at first, then, after 20 classifiers, all scores fluctuated around a number ± 1.5%. Accuracy fluctuated at 88.5%, precision at 88% and recall at 90%, except for MtCE(LSVM-MLP), which was 1% lower. Intuitively, these fluctuations after 20 classifiers mean that adding more classifiers would not improve the measurements a lot. However, since the case is only for fake review detection on Ott et al.'s dataset, we will let the user decide on the number of base classifiers needed for their problem. Next, since we think a 1.5% difference is quite significant, we used the best accuracy results for each MtCE combination for the next batch of experiments. These combinations are 20 classifiers for MtCE(LSVM-MLP), 55 classifiers for MtCE(LR-LSVM-NB), 80 classifiers for MtCE(LR-MLP-DT-NB) and 85 classifiers for MtCE(LR-LSVM-MLP-DT-NB).

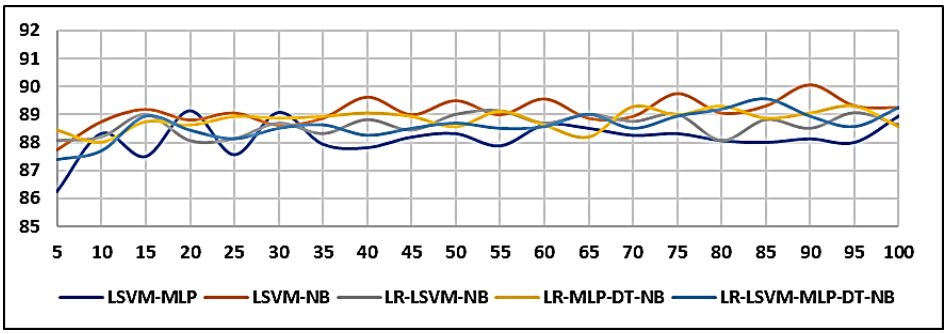

Figure 4: The accuracy (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis)

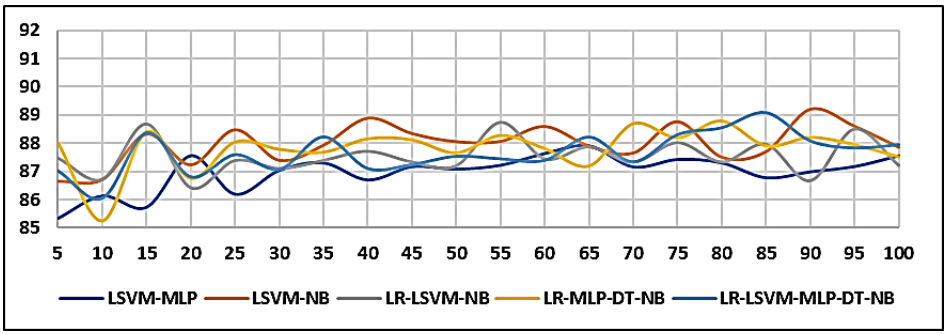

Figure 5: The precision (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis)

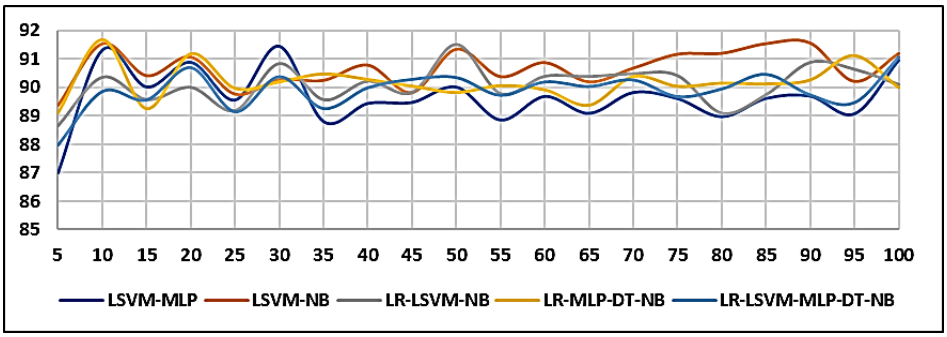

Figure 6: The recall (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis)

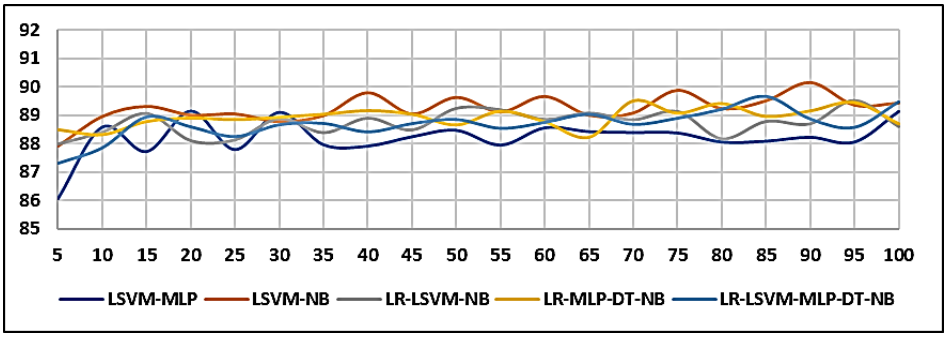

Figure 7: The F-measure (y-axis, in %) of MtCE with 5–100 total base classifiers (x-axis)

5.2.2 EQUALLY SPLIT OR RANDOMLY ASSIGNED BASE CLASSIFIER TYPE

To investigate the equally split vs randomly assigned base classifier type, we ran the experiments 10 times for each equally split and randomly assigned setting for each MtCE combination. The average results and their standard deviation can be seen in Table 6. This table shows that the equally split option gave slightly better average results for all measurements relative to the random option. This implies that when the total of each base classifier is equal, they support each other, and the strength of one type of classifier covers the weakness of the other type when combined in the MtCE. On the other hand, the random option could make one or two base classifiers more numerous than the others, so that they dominated the detection process. The standard deviation of all measurement scores below 1 means the variation of the results from 10 runs was low and close to each other. In more detail, we can see that the standard deviation of 2-combination, the MtCE(LSVM-MLP), on the equally split setting was lower than the random assignment. This is expected since, in an equally split setting, each base classifier total is the same for all 10 runs, while in the random option, this varies. However, the exact opposite happened in 5-combination, the MtCE(LR-LSVM-MLP-DT-NB). Here, the standard deviation of the equally split setting was higher than the random option, implying that the randomness might make the base classifiers more homogenous than split equally between its 5 base classifier types. For the 3- and 4-combinations, the standard deviation scores are similar.

Table 6. Results of equally split vs randomly assigned type of MtCE base classifiers on 10 runs of the 10-fold CV process.

Base Classifier Combination	Equally Split or Randomly Assigned	Accuracy		Precision		Recall		F-measure	
		Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev
LSVM-MLP	Equal	88.21	0.38	86.76	0.35	90.23	0.56	88.39	0.41
	Random	87.82	0.60	86.41	0.59	89.81	0.69	88.00	0.60
LSVM- NB	Equal	89.44	0.36	88.20	0.50	91.12	0.39	89.57	0.37
	Random	89.37	0.19	88.16	0.38	90.93	0.35	89.47	0.22

LR-LSVM-NB	Equal	88.97	0.35	88.14	0.48	90.06	0.42	89.04	0.35
	Random	88.72	0.37	87.79	0.41	89.89	0.49	88.78	0.38
LR-MLP-DT-NB	Equal	89.18	0.32	88.34	0.44	90.24	0.39	89.21	0.29
	Random	89.03	0.35	88.19	0.46	90.12	0.38	89.09	0.35
LR-LSVM-MLP-DT-NB	Equal	88.28	0.66	87.37	0.64	89.59	0.86	88.41	0.67
	Random	88.12	0.37	87.10	0.39	89.49	0.49	88.21	0.39

5.2.3 BOOTSTRAP EFFECT

The subsequent experiments aimed to investigate the bootstrap parameter. We used fixed parameters as above, except the bootstrap setting ranged from 0.25 to 1 (no bootstrapping). As can be seen in Figure 8, the best accuracy, precision and recall were often achieved when the bootstrap setting was 0.5, where every base classifier was trained using 50% of training samples randomly. This indicates that bootstrap setting 0.5 is ideal for the tested MtCE settings, i.e., the five best base-classifier combinations. Therefore, intuitively, it can be used as the default bootstrap setting of MtCE. Based on these findings, for the next set of experiments, we set the bootstrap to be 0.5.

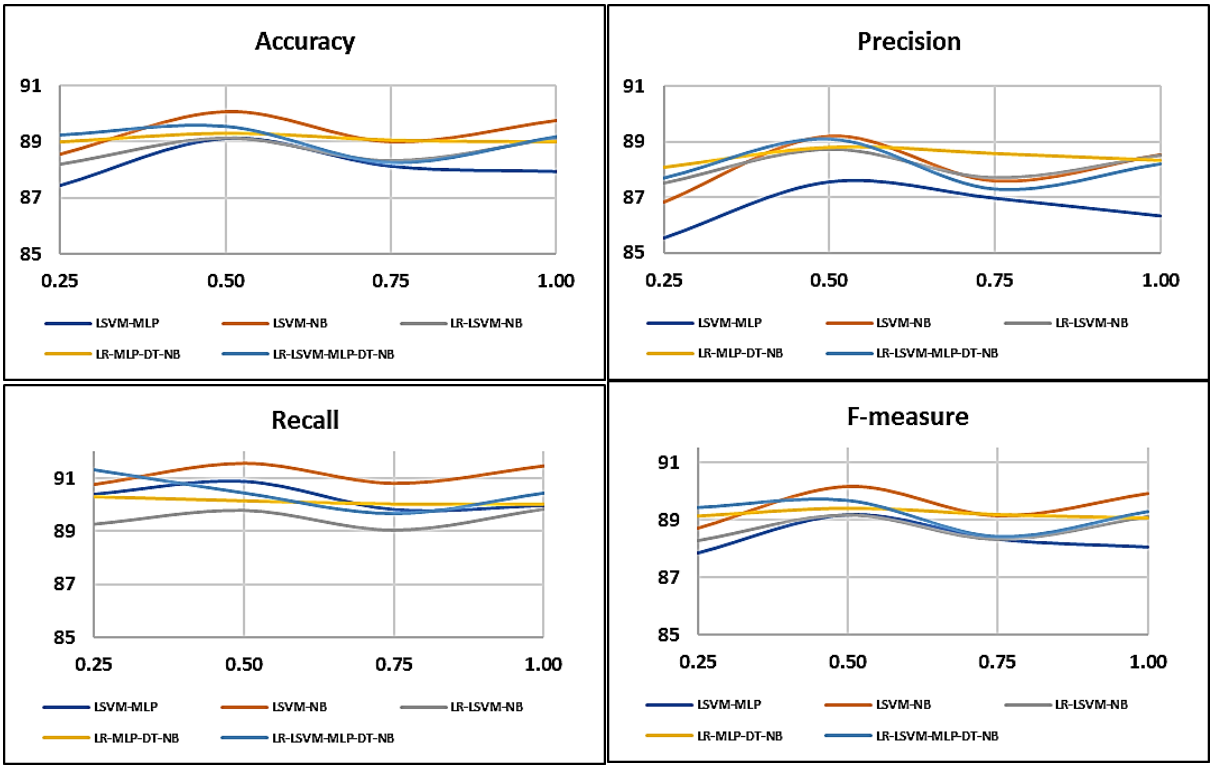

Figure 8: The accuracy, precision, recall and F-measure (y-axis, in %) of the MtCE with bootstrap settings from 0.25 to 1.00 (x-axis)

5.2.4 FINAL OUTPUT: MAJORITY VOTE OR CLASS PRIORITY THRESHOLD

The last parameter to test is how the output of base classifiers should be processed. There are two options in terms of transferring the output of base classifiers to the final output: majority vote and priority class. To test the effect of priority, we set the priority to the positive class (fake class), in the range between absolute priority and 45% priority. As mentioned above, absolute priority means the final result will be recorded as positive/fake if one or more of the base classifiers records this. Percentage priority means the final result will be recorded as positive/fake if the number of base classifiers with this result is the same or more than the percentage threshold. The results of the experiments can be seen in Figure 9.

As per Figure 9, the priority setting increases the recall but decreases other measurements. Moreover, as noted, the higher the recall, the more accurate the positive class (fake class) detection. Thus, we can conclude that the priority setting is helpful when samples of positive class are scarce or when the focus of research is on anomaly detection. The key is how high we choose the percentage of priority so that the increase in recall does not greatly decrease the other measurements. For example, in Figure 9, for 25% priority of MtCE(LSVM-MLP), where recall increased by 5.05%, the accuracy, precision and F-measure decreased by 2.75%, 7.15% and 1.70%, respectively.

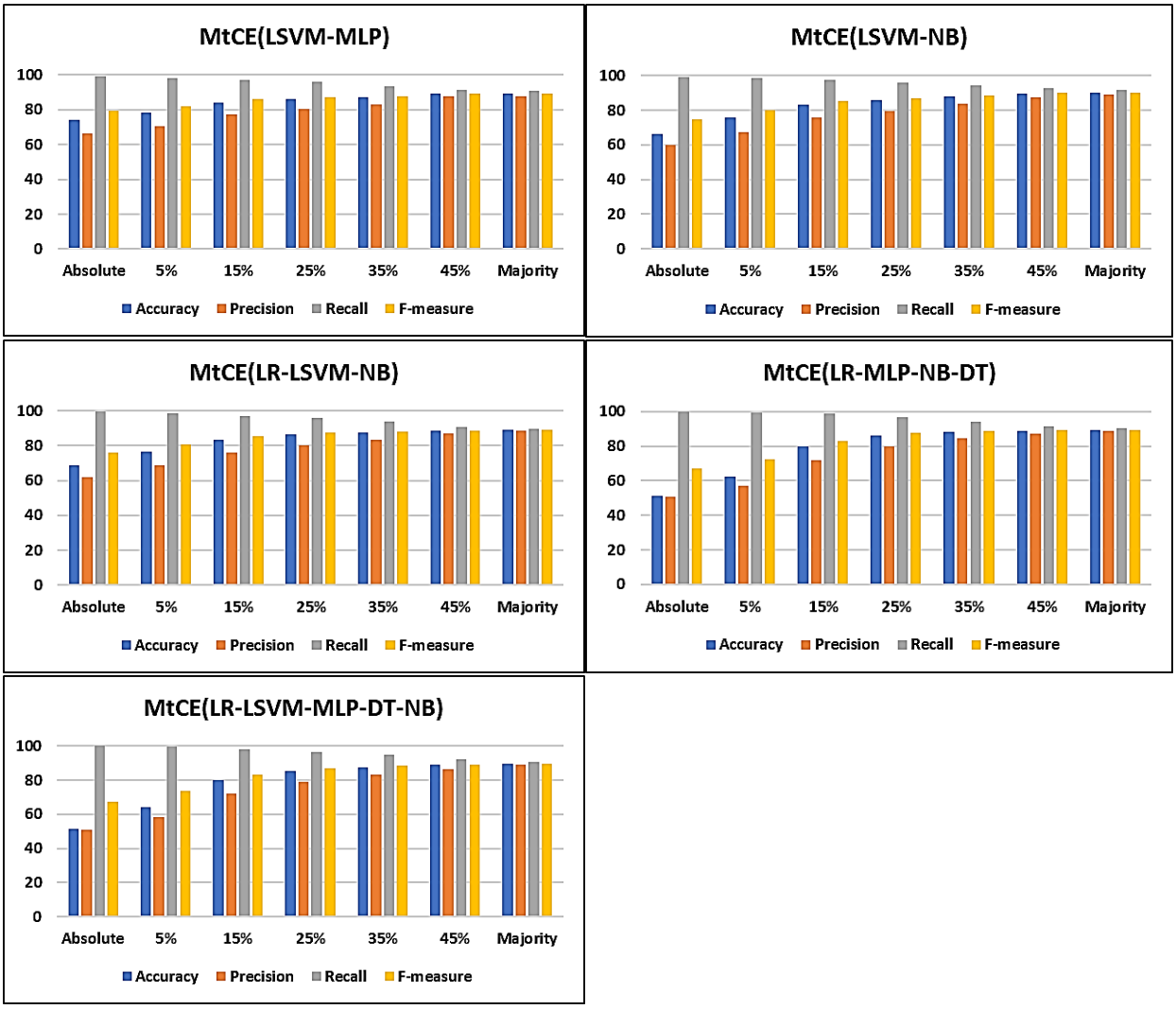

Figure 9: Results of final target experiments, in terms of accuracy, precision, recall and F-measure (y-axis, in %), from absolute priority to 45% priority, and majority (x-axis)

5.3 COMPARISON TO OTHER STUDIES ON THE SAME DATASETS

In this section, we present the MtCE's and other models' best results in light of various considerations. Many factors can influence results, such as the measurement formulas or tools, datasets used, number of records involved in experiments, and how the experiments were conducted (e.g., CV vs traditional train-test-validation split). Therefore, the results presented in Table 7 should be read with the following considerations in mind:

1. We present the MtCE's and other studies' results based on the same datasets (see Table 2). We used all samples that could be processed from each dataset. Note that some studies used only a subset of the dataset

or split the dataset into subsets and measured each subset differently. All of our experiments were conducted using 10-fold CV.

2. It is impossible to ensure that all the studies used the same formulas to measure accuracy, precision, recall and F-measure. Therefore, MtCE results were measured using widely used formulas for binary target prediction (see Table 3). All measurements were in the 'macro' average setting using scikit-learn measurement components [51].
3. The original results from the dataset creators were used as the baseline (see the underlined description and scores in Table 7. We present only the best result for each dataset from each study.

Table 7: Best performance of the MtCE and other studies for the same datasets

No	Dataset	Description	Acc (%)	Pre (%)	Rec (%)	F1 (%)
1	Hotel (various) [49]	Ott et al. [49], SVM on positive sentiment subset	88.4	89.1	87.5	88.3
		Fusilier et al. [27], PU-L modified	-	85.2	72.8	78
		Rout et al. [57], DT	92.11	-	-	-
		Etaiwi and Naymat [21], SVM, all preproc. steps	85	51	86	-
		Zhang et al. [72], DRI-RCNN, 0.9 T.P.	-	-	-	86.59
		Budhi et.al. [12], GB, over-sampling	70.62	67.58	81.29	73.8
		*MtCE(LSVM-NB, 90, Equal, 0.5, Majority)**	*90.06*	***89.19***	***91.56***	***90.16***
2	Hotel (various) [41]	Li et al. [41], OvR SAGE-Unigram	81.8	81.2	84	-
		Ren and Ji [55], Integrated, Employee/Turker subset	92.6	-	-	90.1
		Li et al. [42], SWNN	-	84.1	87	85
		Budhi et.al. [12], GB, under-sampling	71.29	73.61	82.79	77.93
		*MtCE(LR-MLP-DT-NB, 15, Equal, 0.5, Majority)**	*87.82*	***86.67***	***93.26***	*89.81*
3	Restaurant (various) [41]	Li et al. [41], OvR SAGE-Unigram	81.7	84.2	81.6	-
		Ren and Ji [55], Integrated, Customer/Turker subset	86.9	-	-	86.8
		Li et al. [10], SWNN		83.3	88.2	81
		Budhi et.al. [12], GB, over-sampling	72.44	71.23	75.16	73.14
		*MtCE(LR-MLP-DT-NB , 80, Equal, 0.5, Majority)**	*86.07*	***84.47***	***88.89***	*86.37*
4	Doctor (various) [41]	Li et al. [41], OvR SAGE-Unigram	74.5	77.2	70.1	-
		Ren and Ji [55], Integrated, Customer/Turker subset	76	-	-	74.1
		Li et al. [10], SWNN	-	83.7	87.6	82.9
		Budhi et.al. [12], GB, over-sampling	73.35	79.44	86.94	83.02
		*MtCE(LSVM-MLP, 15, Equal, 0.5, Majority)**	***86.56***	***87.05***	***93.12***	***89.88***

* MtCE(<type of base classifiers>, <total base classifiers>, <equally split or random assign of base classifiers>, <bootstrap percentage>, <majority or priority output>)

We do not claim that the MtCE is better or worse than methods in other studies since several factors can affect results. However, we can confidently compare the MtCE to our previous study in fake review detection [12] because the measurement formulas are similar. As is evident in Table 7, compared with our previous approach, the MtCE is superior. Accuracy improves from a minimum of 13.2% in Li et al.'s doctor dataset to a maximum of 19.4% in Ott et al.'s dataset; precision improves from 7.6% to 21.6% and recall from 6.2% to 13.7%.

6 CONCLUSION AND FUTURE WORK

From the experiments above, we can conclude that our proposed ensemble, the MtCE, was able to adequately detect fake or fraudulent reviews, which have become an acute problem on e-commerce platforms. Compared with our previous approach, the MtCE improved accuracy up to 19.4%, precision up to 21.6% and recall up to 13.7%.

Not all combinations of base classifier types produced the expected result of an improvement in the performance on all base classifiers of the MtCE. In some combinations, weak classifiers dragged down the performance of stronger classifiers, especially in terms of accuracy measurements. However, when the combinations were correct, the MtCE

produced better results than its base classifiers in stand-alone modes. It is important to note that while accuracy was not on top, recall was the highest in many cases. This is a positive result since, in binary classification, recall is the same as accuracy of a positive case. In this case, the MtCE was able to better detect the fake review class. Comparison with well-known ensembles, such as the RF, AB, BP and GB, showed that some correct combination of the MtCE could outperform the other ensembles, proving that combining multi-type classifiers in one ensemble process performed better than combining the same classifiers, as is usually done in other ensemble methods.

We proved that DL methods such as the CNN model could be combined with ML counterparts as base classifiers to give better results than if run alone. The number of base classifiers affected performance detection; however, accuracy and other measurements oscillated after 20 base classifiers. While reducing the amount of data for the training process, the bootstrap setting could also affect the performance of the MtCE. From the experiments, we found that 50% bootstrapping gives better results than other percentages, including the non-bootstrap (bootstrap setting 100%). While the majority vote setting for the output offers a fair chance for each class to be detected, the priority threshold boosts detection of the prioritised class. This would be useful if the dataset is imbalanced or the MtCE is used to detect/classify anomalies.

Despite the extensive experimental studies presented in this paper, there remain opportunities/possibilities for further improvement. First, we are aware that multiple types of base classifiers will increase the complexity of the ensemble and increase the cost and time processing for parameter tuning. To overcome this problem, in the near future, we plan to expand on our previously proposed algorithms for ensemble parameter optimisation, namely the Multi-level Particle Swarm Optimisation (PSO) and its parallel version [8]. These nature-inspired algorithms are based on a low-cost PSO algorithm. They were designed to optimise the parameters of an ensemble model, such as BP and AB, choose its base classifier type, and optimise the parameters of these base classifier candidates. Other avenues for future work include investigating the implementation of the MtCE for multi-class problems, heavily imbalanced datasets, and anomaly detection.

REFERENCES

[1] AKRAM, A.U., KHAN, H.U., IQBAL, S., IQBAL, T., MUNIR, E.U., and SHAFI, M., 2018. Finding rotten eggs: A review spam detection model using diverse feature sets. *KSII Transactions on Internet and Information Systems 12*, 10. DOI= http://dx.doi.org/10.3837/tiis.2018.10.026.

[2] ALSUBARI, S.N., SHELKE, M.B., and DESHMUKH, S.N., 2020. Fake reviews identification based on deep computational linguistic features. *International Journal of Advanced Science and Technology 29*, 8s, 3846-3856.

[3] BARBADO, R., ARAQUE, O., and IGLESIAS, C.A., 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management 56*, 4, 1234-1244. DOI= http://dx.doi.org/10.1016/j.ipm.2019.03.002.

[4] BING, L., WONG, T.-L., and LAM, W., 2016. Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. *ACM Transactions on Internet Technology 16*, 2, 1-17. DOI= http://dx.doi.org/10.1145/2857054.

[5] BREIMAN, L., 1996. Bagging predictors. *Machine Learning 24*, 2, 123-140. DOI= http://dx.doi.org/https://doi.org/10.1007/bf00058655.

[6] BREIMAN, L., 1999. Pasting Small Votes for Classification in Large Databases and On-Line. *Machine Learning 36*, 1 (1999/07/01), 85-103. DOI= http://dx.doi.org/10.1023/A:1007563306331.

[7] BREIMAN, L., 2001. Random forests. *Machine Learning 45*, 1, 5-32. DOI= http://dx.doi.org/https://doi.org/10.1023/a:1010933404324.

[8] BUDHI, G.S., CHIONG, R., and DHAKAL, S., 2020. Multi-level particle swarm optimisation and its parallel version for parameter optimisation of ensemble models: a case of sentiment polarity

prediction. *Cluster Computing 23*, 3371-3386. DOI= http://dx.doi.org/https://doi.org/10.1007/s10586-020-03093-3.

[9] BUDHI, G.S., CHIONG, R., PRANATA, I., and HU, Z., 2017. Predicting rating polarity through automatic classification of review texts. In *Proceedings of the 2017 IEEE Conference on Big Data and Analytics (ICBDA)*, Kuching, Malaysia, 19-24. DOI= http://dx.doi.org/https://doi.org/10.1109/ICBDAA.2017.8284101.

[10] BUDHI, G.S., CHIONG, R., PRANATA, I., and HU, Z., 2021. Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis. *Archives of Computational Methods in Engineering 28*, 2543–2566. DOI= http://dx.doi.org/https://doi.org/10.1007/s11831-020-09464-8.

[11] BUDHI, G.S., CHIONG, R., and WANG, Z., 2021. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications 80*, 13079-13097. DOI= http://dx.doi.org/https://doi.org/10.1007/s11042-020-10299-5.

[12] BUDHI, G.S., CHIONG, R., WANG, Z., and DHAKAL, S., 2021. Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. *Electronic Commerce Research and Applications 47, 101048*. DOI= http://dx.doi.org/https://doi.org/10.1016/j.elerap.2021.101048.

[13] CAMPBELL, C. and YING, Y., 2011. *Learning with Support Vector Machines*. Morgan & Claypool.

[14] CARDOSO, E.F., SILVA, R.M., and ALMEIDA, T.A., 2018. Towards automatic filtering of fake reviews. *Neurocomputing 309*, 106-116. DOI= http://dx.doi.org/10.1016/j.neucom.2018.04.074.

[15] CHANG, C.C. and LIN, C.J., 2011. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3, 1-27. DOI= http://dx.doi.org/https://doi.org/10.1145/1961189.1961199.

[16] CHIONG, R., BUDHI, G.S., and DHAKAL, S., 2021. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems 36*, 6. DOI= http://dx.doi.org/https://doi.org/10.1109/MIS.2021.3093660.

[17] CHIONG, R., BUDHI, G.S., DHAKAL, S., and CHIONG, F., 2021. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine 135, 104499*. DOI= http://dx.doi.org/https://doi.org/10.1016/j.compbiomed.2021.104499.

[18] DENG, X., LI, Y., WENG, J., and ZHANG, J., 2019. Feature selection for text classification: A review. *Multimedia Tools and Applications 78*, 3, 3797-3816. DOI= http://dx.doi.org/https://doi.org/10.1007/s11042-018-6083-5.

[19] DONG, M., YAO, L., WANG, X., BENATALLAH, B., HUANG, C., and NING, X., 2018. Opinion fraud detection via neural autoencoder decision forest. *Pattern Recognition Letters*. DOI= http://dx.doi.org/10.1016/j.patrec.2018.07.013.

[20] ELMOGY, A.M., TARIQ, U., and MOHAMMED, A.I.A., 2021. Fake Reviews Detection using Supervised Machine Learning. *International Journal of Advanced Computer Science and Applications 12*, 1, 601-. DOI= http://dx.doi.org/https://dx.doi.org/10.14569/IJACSA.2021.0120169.

[21] ETAIWI, W. and NAYMAT, G., 2017. The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science 113*, 273-279.

[22] FELBERMAYR, A. and NANOPOULOS, A., 2016. The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing 36*, 60-76. DOI= http://dx.doi.org/10.1016/j.intmar.2016.05.004.

[23] FENG, V.W. and HIRST, G., 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of International Joint Conference on Natural Language Processing*, Nagoya, Japan, 338-346.

[24] FREUND, Y. and SCHAPIRE, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences 55*, 1 (1997/08/01/), 119-139. DOI= http://dx.doi.org/https://doi.org/10.1006/jcss.1997.1504.

[25] FRIEDMAN, J.H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 5, 1189-1232.

[26] HAZIM, M., ANUAR, N.B., AB RAZAK, M.F., and ABDULLAH, N.A., 2018. Detecting opinion spams through supervised boosting approach. *PLoS ONE 13*, 6, e0198884. DOI= http://dx.doi.org/10.1371/journal.pone.0198884.

[27] HERNÁNDEZ FUSILIER, D., MONTES-Y-GÓMEZ, M., ROSSO, P., and GUZMÁN CABRERA, R., 2015. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management 51*, 4, 433-443. DOI= http://dx.doi.org/10.1016/j.ipm.2014.11.001.

[28] HEYDARI, A., TAVAKOLI, M., and SALIM, N., 2016. Detection of fake opinions using time series. *Expert Systems with Applications 58*, 83-92. DOI= http://dx.doi.org/10.1016/j.eswa.2016.03.020.

[29] HEYDARI, A., TAVAKOLI, M.A., SALIM, N., and HEYDARI, Z., 2015. Detection of review spam: A survey. *Expert Systems with Applications 42*, 7, 3634-3642. DOI= http://dx.doi.org/10.1016/j.eswa.2014.12.029.

[30] HU, Z., CHIONG, R., PRANATA, I., BAO, Y., and LIN, Y., 2019. Malicious web domain identification using online credibility and performance data by considering the class imbalance issue. *Industrial Management & Data Systems 119*, 3, 676-696. DOI= http://dx.doi.org/https://doi.org/10.1108/IMDS-02-2018-0072.

[31] CHIONG, R., WANG, Z., FAN, Z., and DHAKAL, S., 2022. A fuzzy-based ensemble model for improving malicious web domain identification. *Expert Systems with Applications*. In press.

[32] JACOBS, R.A., JORDAN, M.I., NOWLAN, S.J., and HINTON, G.E., 1991. Adaptive Mixtures of Local Experts. *Neural Computation 3*, 1, 79-87. DOI= http://dx.doi.org/https://doi.org/10.1162/neco.1991.3.1.79.

[33] JAVED, M.S., MAJEED, H., MUJTABA, H., and BEG, M.O., 2021. Fake reviews classification using deep learning ensemble of shallow convolutions. *Journal of Computational Social Science 4*, 2, 883-902. DOI= http://dx.doi.org/10.1007/s42001-021-00114-y.

[34] JIANG, Z., SAINJU, A.M., LI, Y., SHEKHAR, S., and KNIGHT, J., 2019. Spatial ensemble learning for heterogeneous geographic data with class ambiguity. *ACM Transactions on Intelligent Systems and Technology 10*, 4, 1-25. DOI= http://dx.doi.org/10.1145/3337798.

[35] JOHNSON, R.W., 2001. An introduction to the Bootstrap. *Teaching Statistics 23*, 2, 49-54. DOI= http://dx.doi.org/https://doi.org/10.1111/1467-9639.00050.

[36] KERAS, 2019. Keras: The Python Deep Learning library.

[37] KONTSCHIEDER, P., FITERAU, M., CRIMINISI, A., and BUL`O, S.R., 2015. Deep Neural Decision Forests. In *2015 IEEE International Conference on Computer Vision (ICCV)* IEEE, Santiago, Chile, 1467-1475.

[38] KUMAR, A., GOPAL, R.D., SHANKAR, R., and TAN, K.H., 2022. Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems 155*. DOI= http://dx.doi.org/10.1016/j.dss.2021.113728.

[39] KUMAR, N., VENUGOPAL, D., QIU, L., and KUMAR, S., 2018. Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems 35*, 1, 350-380. DOI= http://dx.doi.org/10.1080/07421222.2018.1440758.

[40] LECUN, Y., BOTTOU, L., BENGIO, Y., and HAFFNER, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11, 2278-2324. DOI= http://dx.doi.org/https://doi.org/10.1109/5.726791.

[41] LI, J., OTT, M., CARDIE, C., and HOVY, E., 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, 1566-1576.

[42] LI, L., QIN, B., REN, W., and LIU, T., 2017. Document representation and feature combination for deceptive spam review detection. *Neurocomputing 254*, 33-41. DOI= http://dx.doi.org/10.1016/j.neucom.2016.10.080.

[43] MALBON, J., 2013. Taking fake online consumer reviews seriously. *Journal of Consumer Policy 36*, 2, 139-157. DOI= http://dx.doi.org/10.1007/s10603-012-9216-7.

[44] MANNING, C.D., RAGHAVAN, P., and SCHUETZE, H., 2008. Naïve Bayes text classification. In *Introduction to Information Retrieval* Cambridge University Press, 234-265.

[45] MARTENS, D. and MAALEJ, W., 2019. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering 24*, 6, 3316-3355. DOI= http://dx.doi.org/10.1007/s10664-019-09706-9.

[46] MENARD, S., 2010. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE, Los Angeles.

[47] MOHAMMADI, Y., SALARPOUR, A., and CHOUHY LEBORGNE, R., 2021. Comprehensive strategy for classification of voltage sags source location using optimal feature selection applied to support vector machine and ensemble techniques. *International Journal of Electrical Power & Energy Systems 124*. DOI= http://dx.doi.org/10.1016/j.ijepes.2020.106363.

[48] NORVIG, P., 2016. How to write a spelling corrector.

[49] OTT, M., CARDIE, C., and HANCOCK, J.T., 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, US, 497-501.

[50] OTT, M., CHOI, Y., CARDIE, C., and HANCOCK, J.T., 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings ofthe 49th Annual Meeting ofthe Association for Computational Linguistics*, Portland, Oregon, 309-319.

[51] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., and DUCHESNAY, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 85 (May 08), 2825-2830.

[52] POLIKAR, R., 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine 6*, 3, 21-45. DOI= http://dx.doi.org/10.1109/MCAS.2006.1688199.

[53] QUINLAN, J.R., 1986. Induction of decision trees. *Machine Learning 1*, 1 (March 01), 81-106. DOI= http://dx.doi.org/https://doi.org/10.1007/bf00116251.

[54] REN, Q., LI, M., and HAN, S., 2019. Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. *Big Earth Data 3*, 1, 8-25. DOI= http://dx.doi.org/https://doi.org/10.1080/20964471.2019.1572452.

[55] REN, Y. and JI, D., 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences 385-386*, 213-224. DOI= http://dx.doi.org/10.1016/j.ins.2017.01.015.

[56] REN, Y., ZHANG, L., and SUGANTHAN, P.N., 2016. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. *IEEE Computational Intelligence Magazine 11*, 1, 41-53. DOI= http://dx.doi.org/10.1109/mci.2015.2471235.

[57] ROUT, J.K., SINGH, S., JENA, S.K., and BAKSHI, S., 2016. Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications 76*, 3, 3187-3211. DOI= http://dx.doi.org/10.1007/s11042-016-3819-y.

[58] RUMELHART, D.E., HINTON, G.E., and WILLIAMS, R.J., 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* MIT Press, 318-362.

[59] SAVAGE, D., ZHANG, X., YU, X., CHOU, P., and WANG, Q., 2015. Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications 42*, 22, 8650-8657. DOI= http://dx.doi.org/10.1016/j.eswa.2015.07.019.

[60] SCHAPIRE, R.E., 1990. The strength of weak learnability. *Machine Learning 5*, 2 (1990/06/01), 197-227. DOI= http://dx.doi.org/10.1007/BF00116037.

[61] SHAN, G., ZHOU, L., and ZHANG, D., 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems 144*. DOI= http://dx.doi.org/10.1016/j.dss.2021.113513.

[62] SHIN, J., YOON, S., KIM, Y., KIM, T., GO, B., and CHA, Y., 2021. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics 61*. DOI= http://dx.doi.org/10.1016/j.ecoinf.2020.101202.

[63] SUN, C., DU, Q., and TIAN, G., 2016. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering 2016*, 1-7. DOI= http://dx.doi.org/10.1155/2016/4935792.

[64] TANG, X., QIAN, T., and YOU, Z., 2020. Generating behavior features for cold-start spam review detection with adversarial learning. *Information Sciences 526*, 274-288. DOI= http://dx.doi.org/10.1016/j.ins.2020.03.063.

[65] WAHYUNI, E.D. and DJUNAIDY, A., 2016. Fake review detection from a product review using modified method of iterative computation framework. In *Proceedings of MATEC Web of Conferences 58, 03003*. DOI= http://dx.doi.org/10.1051/matec.

[66] WANG, X., HE, K.L.S., and ZHAO, J., 2016. Learning to Represent Review with Tensor Decomposition for Spam Detection. In *Proceedings ofthe 2016 Conference on Empirical Methods in Natural Language Processing*, Texas, US, 866-875.

[67] WOLPERT, D.H., 1992. Stacked generalization. *Neural Networks 5*, 2 (1992/01/01/), 241-259. DOI= http://dx.doi.org/https://doi.org/10.1016/S0893-6080(05)80023-1.

[68] WU, Y., NGAI, E.W.T., WU, P., and WU, C., 2020. Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems 132*. DOI= http://dx.doi.org/10.1016/j.dss.2020.113280.

[69] XIE, F., FAN, H., LI, Y., JIANG, Z., MENG, R., and BOVIK, A., 2017. Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model. *IEEE Trans Med Imaging 36*, 3 (Mar), 849-858. DOI= http://dx.doi.org/10.1109/TMI.2016.2633551.

[70] XIONG, T., BAO, Y., HU, Z., and CHIONG, R., 2015. Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. *Information Sciences 305*, 77-92. DOI= http://dx.doi.org/https://doi.org/10.1016/j.ins.2015.01.029.

[71] ZHANG, D., ZHOU, L., KEHOE, J.L., and KILIC, I.Y., 2016. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems 33*, 2, 456-481. DOI= http://dx.doi.org/https://doi.org/10.1080/07421222.2016.1205907.

[72] ZHANG, W., DU, Y., YOSHIDA, T., and WANG, Q., 2018. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management 54*, 4, 576-592. DOI= http://dx.doi.org/10.1016/j.ipm.2018.03.007.

5. Second review: Minor revision (22 June 2022)

UNIVERSITAS KRISTEN PETRA

Gregorius S. <greg@petra.ac.id>

FW: ACM TOIT Manuscript TOIT-2021-0307.R1 - decision

1 message

Raymond Chiong <raymond.chiong@newcastle.edu.au> Wed, Jun 22, 2022 at 9:00 AM
To: "Gregorius S." <greg@petra.ac.id>

Ok, this is expected--we are almost there!

Only three points to address:

- However, the explanation for the observation that a weak classifier CNN can significantly improve the overall ensemble performance is still lacking, which could be potentially explained by using ensemble diversity.
- It is mentioned in the related work section that MtCE is different from existing works using heterogeneous models as an ensemble is because MtCE allows any number or type of models to be combined. But what makes existing approach impossible to, e.g., increase the number of models? The reviewer is curious about why the number of models must be fixed in other approaches as it is a hyper parameter to be provided by the user.
- The reviewer is still having a concern on Table 7. Even though the authors mentioned the comparisons across different algorithms are based on unknown and inconsistent experimental setups, it can still mislead the readers. Does this kind of comparisons common in fake news detection literature? If yes, citing several works comparing their approaches in this way can strengthen the results.

Do you have the latest submitted version?

-----Original Message-----
From: Transactions on Internet Technology <onbehalfof@manuscriptcentral.com>
Sent: Wednesday, 22 June 2022 11:51 AM
To: Raymond Chiong <raymond.chiong@newcastle.edu.au>
Cc: Gregorius Satiabudhi <gregorius.satiabudhi@uon.edu.au>; Raymond Chiong <raymond.chiong@newcastle.edu.au>; Arriane.Bustillo@straive.com
Subject: ACM TOIT Manuscript TOIT-2021-0307.R1 - decision

Dear Raymond Chiong,

We have completed the review process for the paper titled, "A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction." has been completed. The decision is to request the authors undergo a minor revision for the manuscript prior before it can be considered further for publication.

Your reviews are listed below. We would suggest that you revise your paper according to the reviewers' comments and resubmit the paper for a second round of reviews. If you wish to revise the paper, your revision due date will be on 12-Jul-2022. Please provide a detailed response to the reviewer comments and indicate (with specificity) which parts of the paper have been modified since the last revision. If you do not intend to submit a revised version of your paper, please let us know so that we can formally close your file. Please be mindful when making your revisions that you still need to maintain the size limitations for papers submitted to TOIT.

Thank you for your submission and we look forward to receiving a future revision. Please feel free to contact the journal administrator Arriane Bustillo at Arriane.Bustillo@straive.com if you have any questions or comments.

Regards,

Prof. Dr. Ling Liu (Georgia Institute of Technology) Editor in Chief, ACM TOIT

Associate Editor's Comments to Author:

Associate Editor
Comments to the Author:
(There are no comments.)

Reviewer(s)' Comments to Author:

Referee: 1

Recommendation: Minor Revision

Comments:
This revised manuscript significantly improved the overall quality of this paper and addressed most of the concerns of this reviewer. In particular, the discussions and Table 1 on related work provide a clear background and justify the motivation of MtCE.

However, the explanation for the observation that a weak classifier CNN can significantly improve the overall ensemble performance is still lacking, which could be potentially explained by using ensemble diversity.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Light

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

Referee: 2

Recommendation: Accept

Comments:
In the revised manuscript, authors included a related work section and emphasized their contributions. My other comments related to the organization and figures are taken into account by the authors and properly refactored.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Light

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

Referee: 3

Recommendation: Minor Revision

Comments:
While the revision addresses some concerns, several questions remain unanswered.

* It is mentioned in the related work section that MtCE is different from existing works using heterogeneous models as an ensemble is because MtCE allows any number or type of models to be combined. But what makes existing approach impossible to, e.g., increase the number of models? The reviewer is curious about why the number of models must be fixed in other approaches as it is a hyper parameter to be provided by the user.
* The reviewer is still having a concern on Table 7. Even though the authors mentioned the comparisons across different algorithms are based on unknown and inconsistent experimental setups, it can still mislead the readers. Does this kind of comparisons common in fake news detection literature? If yes, citing several works comparing their approaches in this way can strengthen the results.

Additional Questions:

Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Moderate

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Maybe
---

6. Second Response to the reviewers' and Second Revision manuscript files are sent to the Correspondent Author (30 June 2022)
    a. Second response to the reviewers'
    b. Second Revision manuscript with brown font for the revision

Dr. Raymond Chiong

Dear Dr. Chiong:

Recently, you received a decision on Manuscript ID TOIT-2021-0307.R1, entitled "A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction." The manuscript and decision letter are located in your Author Center at https://mc.manuscriptcentral.com/toit.

This e-mail is a gentle reminder that your revision will be due shortly. If it is not possible for you to submit your revision by the deadline, please contact Arriane Bustillo at Arriane.Bustillo@straive.com.

Sincerely,

Transactions on Internet Technology

4:41 PM

Yes sir, i am on the way to revise it. I will finish it in 2 days. 7:09 PM

7:20 PM

6/30/2022

Hi Sir, I've sent the 2nd revision of ToiT paper. They are minor revisions so that I cou
finish it quicker.

12:54 P

Reviewer(s)' Comments to Author:

Referee: 1

Recommendation: Minor Revision

Comments:
This revised manuscript significantly improved the overall quality of this paper and addressed most of the concerns of this reviewer. In particular, the discussions and Table 1 on related work provide a clear background and justify the motivation of MtCE.

Response:
Thank you.

However, the explanation for the observation that a weak classifier CNN can significantly improve the overall ensemble performance is still lacking, which could be potentially explained by using ensemble diversity.

Response:
We add a discussion about base classifiers diversity in MtCE would improve the overall performance at the end of paragraph 5 sub-section "5.1. Experimenting With The MtCE For Fake Review Detection"

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Light

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

==========

Referee: 2

Recommendation: Accept

Comments:
In the revised manuscript, authors included a related work section and emphasized their contributions. My other comments related to the organization and figures are taken into account by the authors and properly refactored.

Response:
Thank you for the acceptance and compliment.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Light

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Yes

==========

Referee: 3

Recommendation: Minor Revision

Comments:
While the revision addresses some concerns, several questions remain unanswered.

* It is mentioned in the related work section that MtCE is different from existing works using heterogeneous models as an ensemble is because MtCE allows any number or type of models to be combined. But what makes existing approach impossible to, e.g., increase the number of models? The reviewer is curious about why the number of models must be fixed in other approaches as it is a hyper parameter to be provided by the user.

Response:
We add discussions to answer the curiosity of reviewer in the paragraph 2, and at the end of paragraph 3 of section "2. Related Work On Ensemble Models"

* The reviewer is still having a concern on Table 7. Even though the authors mentioned the comparisons across different algorithms are based on unknown and inconsistent experimental setups, it can still mislead the readers. Does this kind of comparisons common in fake news detection literature? If yes, citing several works comparing their approaches in this way can strengthen the results.

Response:
Presenting the results of previous similar studies in "fake review detection", or machine learning studies in general is common. Of six non-baseline studies that we cited in Table 7, four ([2-5]) presented similar comparisons in their papers. We add citations of these studies at the beginning of the section "5.3 Comparison To Other Studies On The Same Datasets". While Fusilier et al. [1] do not mention it explicitly in their paper, we suspect their baseline measurements are taken from another study. The baseline papers are not comparing their results to other studies because they are the creator of the datasets. Therefore, they are the first.

Additional Questions:
Please help ACM create a more efficient time-to-publication process: Using your best judgment, what amount of copy editing do you think this paper needs?: Moderate

Most ACM journal papers are researcher-oriented. Is this paper of potential interest to developers and engineers?: Maybe
==========

[1] HERNÁNDEZ FUSILIER, D., MONTES-Y-GÓMEZ, M., ROSSO, P., and GUZMÁN CABRERA, R., 2015. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management 51*, 4, 433-443. DOI= http://dx.doi.org/10.1016/j.ipm.2014.11.001.

[2] LI, L., QIN, B., REN, W., and LIU, T., 2017. Document representation and feature combination for deceptive spam review detection. *Neurocomputing 254*, 33-41. DOI= http://dx.doi.org/10.1016/j.neucom.2016.10.080.

[3] REN, Y. and JI, D., 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences 385-386*, 213-224. DOI= http://dx.doi.org/10.1016/j.ins.2017.01.015.

[4] ROUT, J.K., SINGH, S., JENA, S.K., and BAKSHI, S., 2016. Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications 76*, 3, 3187-3211. DOI= http://dx.doi.org/10.1007/s11042-016-3819-y.

[5] ZHANG, W., DU, Y., YOSHIDA, T., and WANG, Q., 2018. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management 54*, 4, 576-592. DOI= http://dx.doi.org/10.1016/j.ipm.2018.03.007.

A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction

Gregorius Satia Budhi*

School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia, Gregorius.Satiabudhi@uon.edu.au

Informatics Department, Petra Christian University, Surabaya 60236, Indonesia, Greg@petra.ac.id

Raymond Chiong*

School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia, Raymond.Chiong@newcastle.edu.au

Abstract: The financial impact of online reviews has prompted some fraudulent sellers to generate fake consumer reviews for either promoting their products or discrediting competing products. In this study, we propose a novel ensemble model—the Multi-type Classifier Ensemble (MtCE)—combined with a textual-based featuring method, which is relatively independent of the system, to detect fake online consumer reviews. Unlike other ensemble models that utilise only the same type of single classifier, our proposed ensemble utilises several customised machine learning classifiers (including deep learning models) as its base classifiers. The results of our experiments show that the MtCE can adequately detect fake reviews, and that it outperforms other single and ensemble methods in terms of accuracy and other measurements in all the relevant public datasets used in this study. Moreover, if set correctly, the parameters of MtCE, such as base-classifier types, the total number of base classifiers, bootstrap and the method to vote on output (e.g., majority or priority), further improve the performance of the proposed ensemble.

Keywords: Fake review detection, online commerce security, novel ensemble model, machine learning, deep learning.

1 INTRODUCTION

Fake review detection is a hot research topic in natural language processing [55]. A fake review is a consumer review of a product with fictitious opinions written deliberately to sound authentic, for commercial motives; that is, to promote or damage the reputation of the reviewed product [42; 50]. There is a significant possibility that fake reviews distort the actual evaluation of a product [23], create harmful effects and eventually reduce trust in consumer reviews and undermine the effectiveness of the online market [43]. These fraudulent acts occur in response to the vital role of consumer reviews in purchasing behaviour in online markets [68]; consumers trust opinions expressed in online reviews and depend on them to make decisions [4; 14; 39]. Without detection efforts, reviews may be replete with lies, fakes and deceptions, and thus completely useless—or worse [55].

The majority of studies in fake review detection follow three main approaches: those based on the content of the reviews, those based on the behaviour of the reviewers, or a combination of these [3; 29]. Content-based methods focus on the linguistic features of text such as words, parts of speech (POS), n-gram, term frequency (TF) or other linguistic characteristics [21; 27; 55]. The behaviour-based approach focuses more on reviewers' behaviour, such as the user's identity, reviewed product, total number of reviews, ratings given, and duration of reviews [1; 3; 59]. This approach is straightforward, needs only a small number of features, and can deliver good results if properly designed. However, the behaviour-based approach is entirely dependent on additional information, such as metadata, provided by the system. In contrast, the content-based approach is independent of the system, requiring only the review text.

* Corresponding authors' emails:
Raymond.Chiong@newcastle.edu.au, Greg@petra.ac.id, Gregorius.Satiabudhi@uon.edu.au

The content-based approach has two sub-categories. The first creates input features from the review text using Bag of Words (BOW), Word2Vec and skip-gram; thus, the input features for detection are words or terms created from the review text itself, and we call this approach textual-based featuring. The features extracted by this approach are more or less similar, mainly in the form of n-gram terms [14; 21; 27; 42; 63]. This textual-based approach is promising and is used in many studies to extract features from review texts. While independent of the system and needing only the text, this approach usually produces good results [11; 21; 27; 41; 42; 49; 55; 63; 72]. The second form of content-based method attempts to extract information and properties from the text, such as the length of the text, total words and total sentences, in addition to linguistic characteristics of the text, such as POS, subjectivity, complexity, diversity and similarity, and sentiment analysis [12; 26; 28; 57; 65; 66; 71]. This type of content-based featuring is lightweight compared with textual-based featuring; for example, only 80 features in [12] compared with 5,000 features in [11]. This approach, too, is independent of the system since it requires only the text; however, performance is usually relatively poor if not combined with other approaches [12].

In this study, we focus on the textual-based approach. For our experiments, we used Ott et al.'s [49] and Li et al.'s [41] datasets. These datasets are public fake review datasets used in many studies (e.g., see [12; 21; 27; 42; 55; 57; 72]). Before extracting the input features, we implemented text preprocessing such as tokenisation, stopword removal, detection of negation words, correction of elongation words, and POS lemmatisation. To extract the input features, we used the BOW method, n-gram words (from unigram to trigram words) and the count vectorised method. These methods convert a collection of text documents to a matrix of token counts. The token count features are in descending order by TF across the corpus (dataset). To detect fake reviews, we propose a novel machine learning (ML) ensemble model—the Multi-type Classifier Ensemble (MtCE)—that utilises several different types of standard (single) classifiers as its base classifier.

Ensemble classifiers are increasingly used in many different areas of study, such as text analysis [10; 39], computer security [30; 31], environmental science [62], medical research [69], electrical fault detection [47] and spatial data processing [34]. The main idea behind ensemble classifier models is to combine multiple single classifiers to produce a combination that outperforms any single classifier itself [8; 56]. In general, ensemble classifier models can be classified into four groups: bagging, boosting, stacked generalisation, and the mixture of experts [52]. Bagging, also called bootstrap aggregating, accumulates multiple predictors/classifiers and runs a plurality vote when predicting classes. These base classifiers are differentiated using a bootstrap technique to replicate the learning set [5]. The Random Forest (RF) and Pasting Small Votes ensemble models are a variation of bagging [6; 7]. The second group of ensemble models is boosting. Similar to bagging, boosting is also an aggregation of multiple classifiers. However, instead of using bootstrap, it uses the boosting technique to train the base classifiers. This technique can convert weak learning algorithms to achieve arbitrarily high accuracy [60]. Adaptive Boosting (AB) [24] and Gradient-Boosted Trees (GB) [25] are two well-known algorithms that use this technique. In stacked generalisation [67], the classifiers are stacked to one another so that the output of some first-level base classifiers becomes the input of a second-level meta classifier. The combination of expert methods [32] combines the weighted output of several base classifiers in the consecutive combiner.

While promising, we see the potential for improvement from these ensemble classifiers; they typically use only the same type of single classifier as the base classifier. To differentiate the output of each base classifier, they implement bootstrapping or boosting of the training sample inputs via bagging or boosting, stacking the classifiers, or applying different weights for their output. In contrast, our proposed MtCE combines several types of single classifiers working together as an ensemble model. As every single classifier has strengths and weaknesses in classification or detection, by combining different types of single classifiers in an ensemble, we use the strength of one classifier to 'cover' for the weakness of the other classifiers, and vice versa. The types of base classifiers ensembled in our MtCE can be customised by the user based on the problem and their experience and knowledge of this problem. To further increase the performance of our model, we implemented the bootstrap technique [5] and the method to vote on output (majority or priority) to produce the final output. For the base classifiers of our proposed ensemble model, we implemented several standard single classifiers that performed well in previous research on text mining [9-12; 16; 17], including the Logistic Regression (LR) [46], Linear-kernel Support Vector Machine (LSVM) [13; 15], Multi-Layer Perceptron (MLP)

[58], and Convolutional Neural Network (CNN) [40]. In addition, two other classifiers that are usually applied as the base classifiers in several ensemble models, namely Decision Tree (DT) [53] and Naïve Bayes (NB) [44], were also used as base classifiers. The MtCE was compared with ensemble models widely used in the literature, including Bagging Predictors (BP) [5], RF, AB and GB.

Theoretically, our work contributes to the artificial intelligence research domain, especially ML, by proposing a novel ensemble approach that can solve classification, detection and prediction problems. Most ensemble models in the literature either (1) have only one type of base classifier (such as RF, AB, BP and GB) or (2) have a small combination of fixed types and number of base classifiers that are designed to solve a particular problem [33; 55; 63; 64]. In contrast, our proposed model (the MtCE) provides the freedom to select the types and number of base classifiers depending on one's requirements. Bootstrapping is included in the MtCE process to improve the model's stability and accuracy. Finally, a priority threshold approach is introduced, in addition to majority voting, to decide the final result; this priority threshold approach can improve the recall and overcome the imbalanced issue.

From a practical perspective, this work also contributes to addressing online commerce security problems—we have successfully implemented a model that can detect fake online consumer reviews with better results than previous research. Product reviews from buyers are commonly used as a guide by most consumers to make purchasing decisions on electronic commerce these days [9]. Reliable and effective detection of fake reviews will increase the trustworthiness of online commerce[10]. On the other hand, sellers use consumer reviews to evaluate the brand perception and consumers' satisfaction [22]. For this reason, maintaining the quality of product reviews is vital. Therefore, fake review detection is an urgent task and a high priority for online commerce portal providers.

The rest of this paper is organised as follows. In the next section, we explain the design of our proposed model, the MtCE. After that, we discuss how we conducted experiments to test the performance of the MtCE, such as the datasets we used, testing framework, details of the textual-based approach, types of base classifiers, and the measurements. In Section 4, we discuss the results and analyses, and finally, we draw conclusions and outline the possibilities for future work.

2 RELATED WORK ON ENSEMBLE MODELS

The previous section provided a concise introduction to both fake review detection and ensemble models. This section discusses recent fake review detection studies from the literature (2016 to the present) that have proposed new ensemble models or implemented existing ensemble models for detection of fake reviews (see Table 1).

The objective of most previous studies has been to propose feature extraction approaches for the input of standard ML and deep learning (DL), including their ensemble models. As discussed in Section 1, these feature extraction approaches can either be textual-based, which creates features from the words/terms of the review text itself [11; 20; 21; 72]; or content-based, which creates features from the text and extract features from the text's information, properties, sentiment polarities, and characteristics [2]; or behaviour-based, which emphasises the behaviour of the reviewers rather than their reviews [3; 19; 39; 45].

The above approaches have also been combined to improve the detection results [12; 26; 38; 61; 71]. These studies investigated existing ML, DL and ensemble models (i.e. AB, BP, RF, and GB) to detect fake reviews using features proposed using the combined approaches. The base classifiers of these ensemble models are one type; for example, the RF utilises decision trees for its base classifiers, and GB utilises regression trees. Both ensembles utilise and modify the trees in their processes. Therefore, it is impossible to replace their tree-type base classifiers with others. While the type of the base classifier in some ensemble methods can be changed, by design, the replacement must still be one type, i.e., it is impossible to mix more than one type of base classifier for AB and BP.

Some studies in the literature have also proposed novel ensemble models. For example, Sun et al. [63] proposed a bagging style of 2 SVMs and a CNN to detect fake reviews. They utilised special SVMs, namely $BIGRAMS_{SVM}$ and $TRIGRAMS_{SVM,}$ to detect terms features from the review text input, with a unique CNN model (Product Word Composition Classifier (PWCC) to capture product-related review features. The output of these three classifiers were

processed in a bagging style method to produce the final detection result. Similarly, a forest of decision trees, called the Neural Autoencoder Decision Forest, was proposed by Dong et al. [19] to detect fake reviews. Their ensemble was designed by combining the Deep Neural Decision Tree (proposed by Kontschieder et al. [37]) with the Autoencoder method. An integration of several CNNs and Gated Recurrent Neural Networks (GRNNs) were introduced by Ren and Ji [55] in 2017. The CNNs were implemented to produce continuous sentence vectors from words, then handled by several forward and backward GRNNs to produce document representations of the reviews. A unique model of CNN, namely bfGAN, was proposed by Tang et al. [64] in 2020, and combined with SVM for fake review detection. The authors claim that their model can effectively detect cold-start spam/fake reviews; the cold start problem is when a new reviewer posts a review. Javed et al. [33] proposed an ensemble of three classifiers (CNNs) for fake review detection. Each classifier detects different feature groups: a textual-based, a content-based (non-textual), and a behaviour-based group of features; a voting mechanism is the used to produce the final detection. The above-discussed approaches are similar in terms of one important aspect: they have proposed a specific ensemble model with specific types and total number of base classifiers. In this kind of ensemble models, the types and numbers of their base classifiers could not been modified since each base classifier has a specific task and purpose.

Our proposed ensemble model, the MtCE, is different. In the MtCE, diverse base classifiers can be mixed together so that the strength of one type of base classifier can overcome the weaknesses of other types of base classifier and vice versa. The total number of base classifiers and the number of each type of base classifier can also be configured freely based on specific requirements. Our model also applies the bootstrapping technique to ensure stability and improve accuracy. For the final classification result, we introduce a priority threshold technique that prioritises a particular class in conjunction with common majority voting that is usually applied by other ensembles. This priority threshold technique can overcome the imbalanced problem when applied to the scarce/rare class as well as improve the recall in binary classification if applied to the main (positive) class.

Table 1. An overview of related work (those algorithms in italic are ensemble models)

Author	**Year**	**Featuring type**	**Algorithm**
Sun et al. [63]	2016	Textual-based	*Bagging of 2 SVMs and PWCC*
Zhang et al. [71]	2016	Content- and behaviour-based	NB, SVM, DT, *RF*
Etaiwi and Naymat [21]	2017	Textual-based	NB, SVM, DT, *RF, GB*
Ren and Ji [55]	2017	Textual-based	SVM, GRNN, Recurrent Neural Network (RNN), CNN, *CNN-GRNN (Integrated)*
Dong et al. [19]	2018	Behaviour-based	*Neural Autoencoder Decision Forest*
Hazim et al.[26]	2018	Content and behaviour-based	*Generalised Boosted Regression Model (GBM) Gaussian, GBM Poisson, GBM Bernoulli, GB, AB*
Kumar et al. [39]	2018	Behaviour-based	LR, k-NN, NB, SVM, *RF, AB*
Zhang et al. [72]	2018	Textual-based	Recurrent CNN, SVM, CNN, *CNN-GRNN*
Barbado et al. [3]	2019	Behaviour-based	LR, DT, Gaussian NB (GNB), *RF, AB*
Martens & Maalej[45]	2019	Behaviour-based	DT, MLP, LSVM, RBF kernel SVM (RSVM), GNB, *RF*
Tang et al. [64]	2020	Behaviour-based	CNN, *CNN-bfGAN + SVM*
Alsubari et al. [2]	2020	Content-based	DT, *RF, AB*
Budhi et al. [11]	2021	Textual-based	LR, MLP, LSVM, *RF, AB, BP*
Budhi et al. [12]	2021	Content- and behaviour-based	DT, LR, LSVM, MLP, CNN, *RF, GB, AB, BP*
Elmogy et al. [20]	2021	Textual-based	LR, NB, k-NN, SVM, *RF*
Javed et al. [33]	2021	Textual-, content-, and behaviour-based	*Ensemble of 3 CNNs*
Shan et al. [61]	2021	Content- and behaviour-based	DT, SVM, NB, MLP, *RF*
Kumar et al. [38]	2022	Textual-, content-, and behaviour-based	Long short-term memory, Artificial Neural Network, RNN, RSVM, LR, k-NN, NB, *RF, GB, Light GB Method*

3 THE MTCE

The MtCE is a novel ensemble of classifiers developed in light of the fact that every single classifier has its strengths and weaknesses (see Figure 1). This ensemble is proposed based on the following idea:

> *Every single classifier has been created with a different method, formulation and purpose, and thus, has strengths and weaknesses in different spots; for example, (1) for the same dataset, each classifier may correctly classify the different set of records; (2) a classifier is usually created to solve one or two problems, and not tested for the case of other problems; and (3) some classifiers perform well for smaller datasets but not for bigger ones, or vice versa. Therefore, by combining different single classifiers in one ensemble, the strengths of one classifier may 'cover' for the weaknesses of another.*

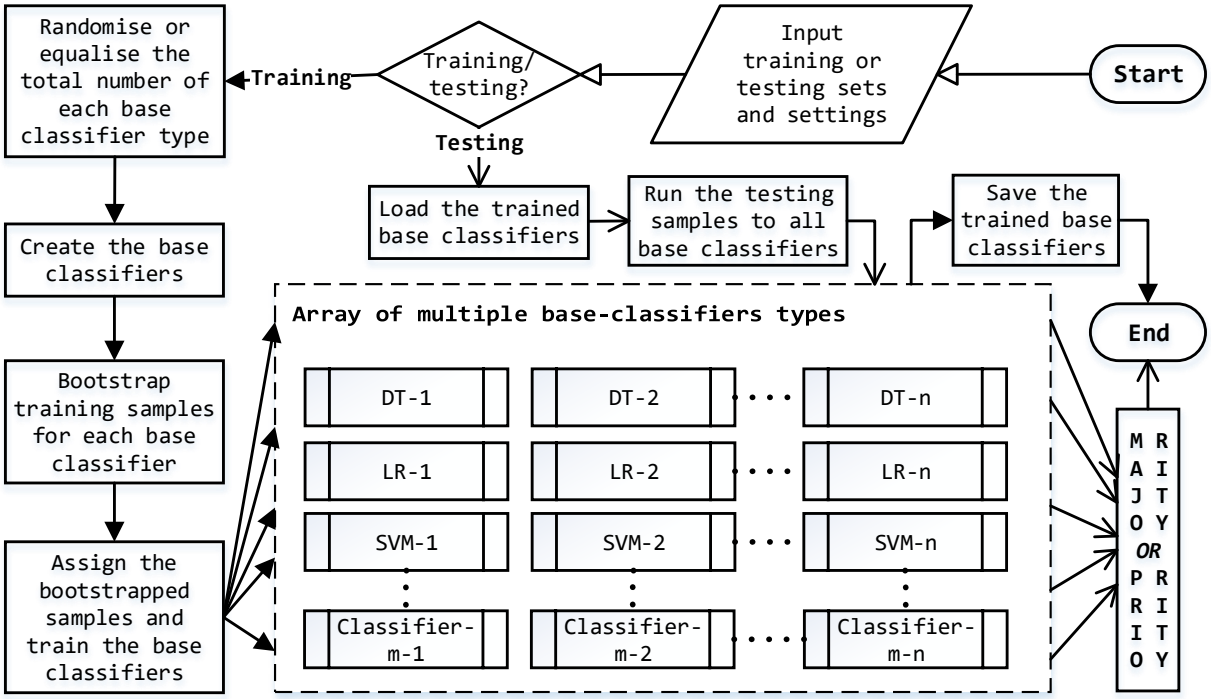

Figure 1: Design of the MtCE

In Figure 1, we depict two processes: the training process (the full head arrows ➔) and the testing process (the line head arrows →). All processes begin with inputting the training or testing sets and the parameter settings of the MtCE (see hollow head arrows -▷).

3.1 BASE CLASSIFIERS OF THE MTCE

Base classifiers here include several classifier objects that are utilised within an ensemble model. These classifiers will be run jointly in a specific way (depending on the ensemble design) to produce results. These results are then processed to produce the final result of the ensemble model. For our MtCE, first the user will decide on the types of base classifiers and their total numbers. After that, for the training process, the user determines if the base classifiers will be created in equally split numbers between the types setting or in random fashion. For example, suppose the total base classifiers are 10 of 3 types (LR, DT, NB); an equally split setting will create 4 LR, 3 DT and 3 NB, while for a random setting, each classifier type will be randomised from 1 to 8 (total classifiers – (total types – 1)) classifiers. While splitting the number of base classifier types equally is more sensible if the user does not know the exact nature of the problem at hand, randomising them sometimes could give better results. Moreover, with a simple modification in the implementation phase, the user can also set the exact different total number of each base classifier type, supposed

they have a reason to do so. In this study, we did not test the effect of setting exact numbers of each type of base classifier since, for our problem, it will become similar to the randomised setting.

3.2 BOOTSTRAP SAMPLING

After creating the base classifiers, each base classifier is assigned a subset of training samples based on the bootstrap sampling process. Bootstrapping the samples could improve the stability and accuracy of the model [5]. The bootstrap sampling method draws sample data repeatedly with replacement from the source (data), based on the percentage setting [5; 35]. After the training process for each base classifier finishes, the weights and parameters of all base classifiers are saved in a file.

3.3 DETERMINING THE FINAL OUTPUT

The testing process is straightforward. After inputting the testing samples and loading the previously trained base classifiers, all samples are run with the base classifiers. The final output will be determined by one of two processes: the majority vote or the priority class threshold. The majority vote chooses the majority class outputs from the results of the base classifiers. If the majority votes of all classes are equal, the final result is the smallest class in the corpus. The priority class threshold provides a final result based on the priority class chosen in the setting; that is, if the positive class is chosen as the priority, then, if the positive class outputs from base classifiers are the same as or more than the threshold, the final output is a positive class. The priority class threshold can be set from absolute, which needs only one base classifier to pick the chosen class, to the maximum threshold before it becomes a majority vote (i.e. more than 50% + 1 in a binary classification problem). This priority class threshold mechanism will focus on the class that is prioritised and will increase the performance of detection of this class.

4 TESTING THE PERFORMANCE OF THE MTCE

4.1 FRAMEWORK FOR TESTING

To evaluate our proposed ensemble, the MtCE, we implemented the comparison framework that we used previously to investigate classifiers for sentiment polarity detection of consumer reviews [10]. In this study, we modified the framework so that it could be applied to test the MtCE for fake review detection (see Figure 2). We used this framework to test different settings of the MtCE using similar datasets and comparing their performances. We ran the same test using several single classifiers as base classifiers of the MtCE and several well-known ensembles for comparison.

As can be seen in Figure 2, after a review text is loaded to a string, it is processed in the preprocessing subroutine to remove unnecessary words and corrections. Features are then extracted from the texts using a textual-based approach, as discussed in Section 1. After combining with the designated targets, these feature-target vectors are split into 10 folds for the cross-validation process. This same 10-fold setting is then trained and tested on several different combinations of MtCE models. Afterwards, the measurement results of these different MtCE models are compared to determine the best setting of the MtCE for fake review detection.

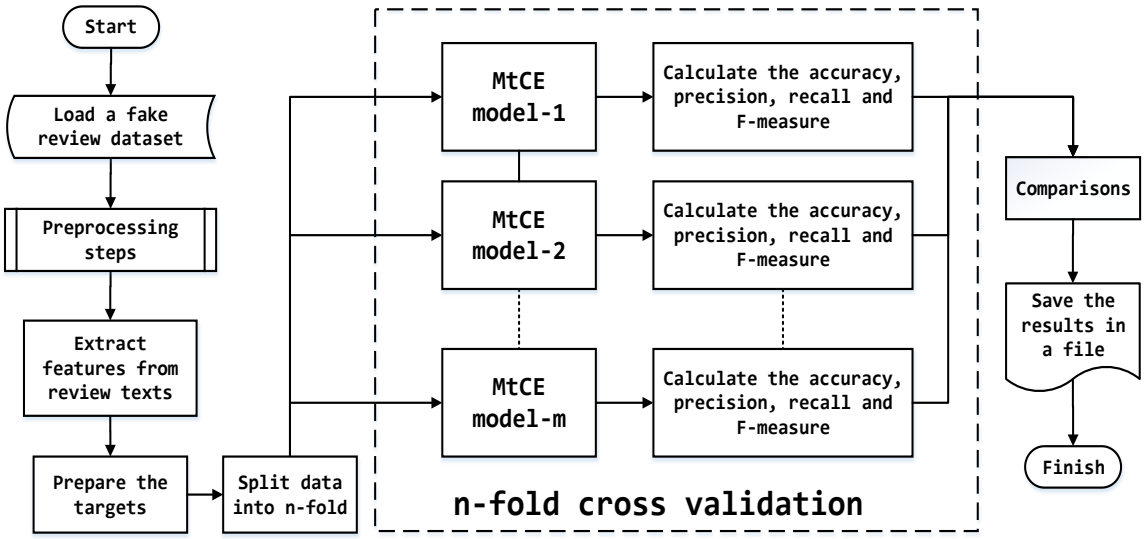

Figure 2: Framework to test and compare the MtCE

In the preprocessing steps (see Figure 3), we implemented several methods as follows:

a. removal of punctuation, numbers and stopwords
b. correction of spelling errors, elongation words and negative words
c. POS tagging and lemmatisation.

Punctuation and number removal is a necessary step for a textual-based approach since the input features of this approach are the words. Therefore, we needed to remove the non-word parts of the text reviews. Stopword is a term for words commonly used in English sentences such as ‘the’, ‘is’, ‘at’, ‘which’ and ‘on’. These words can be removed from the text before it is used as a training record. These kinds of words may mislead detectors in recognising the trait words of a class because they are often the most common words in a corpus.

Misspelled words create unnecessary issues for detection since the detector will recognise them as different words from their correct forms. Like misspelled words, elongation words such as ‘yesss’, ‘fiiine’ and ‘yoouu’ can also increase the diversity of words in the training example and make the classifiers harder to train. We utilised Peter Norvig’s code for spelling correction [48] to correct misspelled and elongation words. Negative words have many forms, depending on the grammar, but their purpose is similar—to change a positive sentence to be negative. Therefore, we shifted all negative words to their primary form to reduce the diversity of words.

POS tagging categorises words by their syntactic function, such as noun, pronoun, adjective, verb and adverb. This step is essential because it puts the words in their context [21] so that in the lemmatisation process, we can lemmatise the word to the correct context. Lemmatisation is a method for changing the word back to its basic form; POS lemmatisation returns the chosen word to its basic syntactic function. This step reduces the diversity of words inside the dataset and makes it easier to recognise.

Because we extracted the input features directly from social media messages, we used the BOW feature extraction method commonly used for textual-based features [18]. This method works by decomposing the entire text into a group of singular words. Similar to previous work [11], in this study, we used it in combination with the n-gram technique to capture singular words and terms that convey one meaning but are composed of multiple words. For terms, the BOW checks for the existence of a contiguous sequence of n words from the given text sample. This study limited this checking to trigrams since sequences of more than three words are rare in real-world texts. After decomposing the text, the terms were sorted based on their frequency across the corpus.

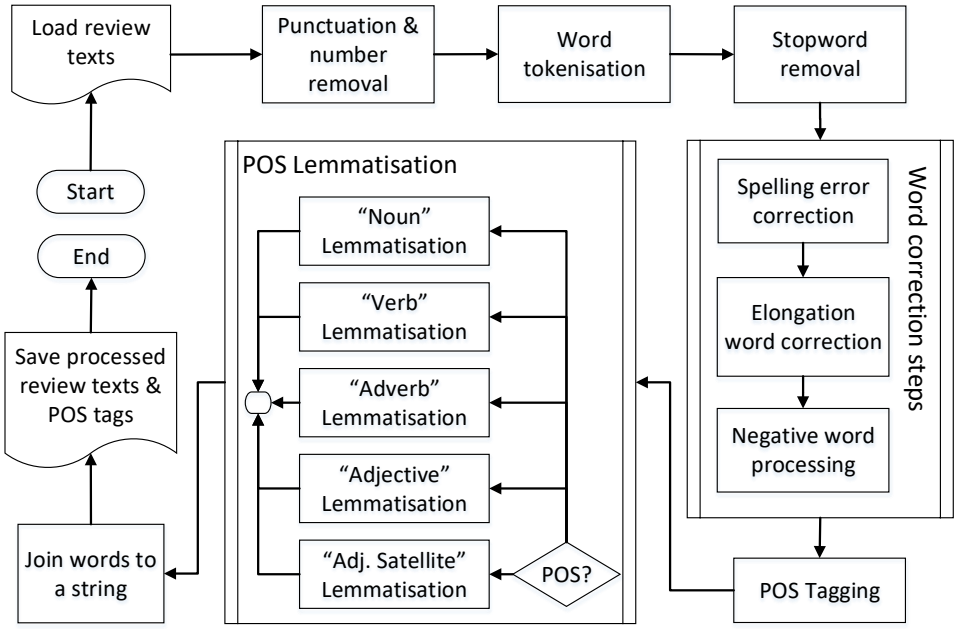

Figure 3: Preprocessing steps [11]

4.2 BASE CLASSIFIERS AND ENSEMBLE CLASSIFIERS FOR EXPERIMENTAL PURPOSE

As mentioned in Section 2, our proposed ensemble can apply multiple types of single classifiers. While in theory, any single classifier could be used as the base classifier of the MtCE, to test its performance, we investigated several combinations of ML and DL that performed well in previous research on text mining [9-12; 16; 17], which include the LR, LSVM, MLP, and CNN models. We also investigated two other classifiers, DT and NB, since these two models are commonly used as default base classifiers in some ensemble models, such as the RF [7], BP [5] and AB [24]. For comparison purposes, we tested some ensemble models BP, RF, AB and GB [25].

We built all ML classifiers and ensembles, except the CNN, using scikit-learn components [51]; the CNN was built with Keras [36] wrapped with scikit-learn components (scikit-learn does not provide a CNN component but a way to wrap DL components of Keras to be used with or within scikit-learn). To ensure the results could only be affected by our model implementation and not by modifying classifier parameters, we used the default parameters for all single classifiers used as base classifiers and the ensemble models.

4.3 DATASETS

Intuitively, our proposed ensemble classifier can be applied to a wide range of problems, similar to other ML ensemble classifiers. However, to test the performance of the MtCE, we conducted experiments using four public fake review datasets (see Table 2 for additional details). The public datasets from Ott et al. [49] and Li et al. [41] were created using domain experts, such as Amazon's Mechanical Turk service (Turkers) and hotel employees, to provide false/fake reviews for some online services. Li et al.'s hotel dataset is an extension of Ott et al.'s dataset.

Table 2: Statistics for several public datasets

Dataset Author	Domain	Total Records	Fake		Genuine	
			Total	%	Total	%
Ott et al. [49]	Hotel (Various)	1600	800	50.00	800	50.00
Li et al. [41]	Doctor (Various)	558	357	63.98	201	36.02
	Hotel (Various)	1880	1080	57.45	800	42.55
	Restaurant (Various)	402	202	50.25	200	49.75

4.4 CROSS-VALIDATION AND MEASUREMENTS

We evaluated the performance of the MtCE using n-fold Cross-Validation (CV), which is widely applied to evaluate the generalisation performance of an algorithm [30], to reduce bias between the dataset and the training or testing set [54], and avoid overfitting [70]. We implemented scikit-learn [51] for performance measurements of our experiments. These measurements and their respective formulas can be found in Table 3.

Table 3: Measurement functions and formulas

No	Name	Sklearn Function	Equation
1	Accuracy	accuracy_score()	$A(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{k=0}^{n_{samples}-1} 1(\widehat{y_k} = y_k)$ where y is the set of predicted pairs, $\hat{y}$ is the set of true pairs and $n_{samples}$ is total samples.
2	Precision	precision_score()	$P(y_k, \widehat{y_k}) = \frac{tp}{tp + fp}$ where k is the set of classes, y_k is the subset of y with class k, tp is true positive, and fp is false positive.
3	Recall	recall_score()	$R(y_k, \widehat{y_k}) = \frac{tp}{tp + fn}$ where fn is false negative.
4	F-measure/F1	f1_score()	$F_1(y_k, \widehat{y_k}) = 2 * \frac{P(y_k, \widehat{y_k}) * R(y_k, \widehat{y_k})}{P(y_k, \widehat{y_k}) + R(y_k, \widehat{y_k})}$

5 RESULTS AND DISCUSSION

5.1 EXPERIMENTING WITH THE MTCE FOR FAKE REVIEW DETECTION

The first group of experiments aimed to test the performance of the MtCE for fake review detection. For the base classifiers, we implemented CNN, LR, LSVM and MLP, which were performing well in our previous studies [9-12]. We also implemented DT and NB, which were used as default base classifiers in many ensembles, such as BP, AB and RF. These base classifier parameters are the defaults of the scikit-learn components used to implement these classifiers [51]. As shown in the Appendix, we tested all possibilities from 2-, 3-, 4- to 5-combination and presented 11 selected combinations in Table 4. Besides combining base classifiers, the other settings were fixed: total base classifiers = 15, shared equally for each type; bootstrap = 0.5; and final output = majority vote. The datasets we used in the experiments were Ott et al.'s dataset [49] and all three of Li et al.'s datasets [41]. All experiments were conducted using the 10-fold CV method. For the comparison, we also predicted the datasets using single classifiers as above and several well-known ensemble classifiers (RF, BP, AB and GB). The results of the single and ensemble classifiers are provided in Table 5.

Table 4: Results of the combination of base classifiers for the MtCE model (selected)

Num[1]	MtCE Base Classifier Combination	Ott et al.'s Dataset[2]				Li et al.'s Dataset (Doctor)[2]			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
2	LR-MLP	**88.00**	*86.85*	**89.44**	**88.08**	84.41	*84.23*	**92.97**	88.33
5	LSVM-MLP	87.50	85.72	**90.02**	**87.74**	**86.56**	**87.05**	**93.12**	**89.88**

7	LSVM-NB	**89.19**	**88.32**	**90.43**	**89.31**	*84.75*	86.85	**90.32**	88.19
9	MLP-NB	**89.63**	**88.52**	**91.22**	**89.80**	*85.31*	*84.00*	**95.30**	*89.12*
18	LR-LSVM-NB	**89.00**	**88.68**	**89.59**	**89.07**	84.58	85.70	**91.44**	88.34
22	LSVM-MLP-NB	**89.13**	87.86	**91.00**	**89.31**	85.13	86.39	91.88	88.79
26	CNN-LR-MLP	**88.19**	**87.32**	**89.33**	**88.27**	84.40	83.75	**93.51**	88.29
35	LR-LSVM-MLP-NB	**89.19**	**88.77**	**89.95**	**89.28**	85.66	87.01	91.38	88.94
37	LR-MLP-DT-NB	**88.75**	**88.41**	89.25	**88.79**	84.06	83.85	**93.06**	88.04
47	LR-LSVM-MLP-DT-NB	**88.94**	**88.38**	**89.56**	**88.94**	84.24	84.89	92.00	88.12
50	CNN-LR-LSVM-MLP-NB	**88.88**	**88.21**	**89.83**	**88.95**	86.21	86.45	**93.18**	89.54
		Li et al.'s Dataset (Hotel)[2]				Li et al.'s Dataset (Restaurant)[2]			
2	LR-MLP	**87.34**	85.09	**94.30**	**89.45**	85.11	83.99	**88.36**	85.69
5	LSVM-MLP	86.12	84.71	**92.67**	**88.44**	**85.83**	83.98	**88.55**	**86.02**
7	LSVM-NB	**87.45**	87.09	**91.75**	**89.34**	85.35	84.51	86.33	85.21
9	MLP-NB	87.23	86.69	**91.92**	**89.17**	86.04	85.36	87.18	**86.15**
18	LR-LSVM-NB	85.96	85.20	**91.34**	88.14	**86.58**	85.71	**88.08**	**86.41**
22	LSVM-MLP-NB	86.97	86.09	**92.34**	**89.05**	85.32	83.68	**88.65**	85.75
26	CNN-LR-MLP	85.85	84.79	**91.96**	88.17	83.34	82.42	84.09	82.88
35	LR-LSVM-MLP-NB	86.70	85.12	**92.92**	88.78	82.32	81.51	84.77	82.80
37	LR-MLP-DT-NB	**87.82**	86.67	**93.26**	**89.81**	84.38	83.61	85.67	84.57
47	LR-LSVM-MLP-DT-NB	**87.55**	86.19	**93.28**	**89.58**	84.63	83.29	87.36	85.06
50	CNN-LR-LSVM-MLP-NB	86.76	85.98	**91.98**	**88.81**	83.80	82.59	86.17	83.98

[1] The number here is associated with the number in Appendix
[2] **Bold-green** = the MtCE result is higher than that of all base classifiers in Table 5; *Italic-red* = the MtCE result is higher than all base classifiers; Normal-black = at least one base classifier result is higher than that of the MtCE.

Table 5: Results of single and ensemble classifiers

Num.	Classifiers	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	CNN	79.63	81.48	77.36	79.13	67.76	70.64	88.40	77.02
2	LR	87.13	87.05	87.25	87.11	83.35	85.48	89.71	87.26
3	LSVM	85.88	85.67	86.21	85.89	83.13	86.28	87.80	86.91
4	MLP	87.56	87.05	88.37	87.66	85.67	85.89	92.94	89.16
5	DT	71.50	72.21	69.71	70.83	65.79	72.98	74.01	73.32
6	NB	88.50	87.97	89.45	88.63	87.12	90.68	88.98	89.68
7	RF	87.00	87.44	86.62	86.92	76.35	74.05	97.24	83.98
8	BP	75.00	74.23	76.79	75.41	74.19	74.50	90.21	81.45
9	AB	80.06	79.64	80.82	80.09	74.93	79.56	81.28	80.16
10	GB	82.81	83.67	81.63	82.56	76.52	77.08	90.63	82.99
		Li et al.'s Dataset (Hotel)				Li et al.'s Dataset (Restaurant)			
1	CNN	78.83	82.48	80.63	81.29	71.63	71.66	73.20	71.75
2	LR	85.59	85.50	90.25	87.77	81.59	81.03	84.52	81.74
3	LSVM	84.20	84.59	88.26	86.33	82.11	79.93	84.40	81.53
4	MLP	86.12	85.89	90.82	88.24	85.56	84.85	87.69	85.73

5	DT	68.88	73.81	71.15	72.38	60.24	60.34	63.27	61.11
6	NB	87.29	89.73	87.92	88.78	86.08	86.36	86.63	86.10
7	RF	81.97	80.78	90.28	85.17	81.34	78.31	87.86	81.88
8	BP	75.96	76.35	84.14	80.03	71.65	71.08	74.51	71.95
9	AB	79.89	80.77	84.87	82.67	70.89	70.64	70.35	70.38
10	GB	80.48	78.83	89.94	84.00	76.12	77.43	75.46	75.78

Comparing Table 4 with Table 5, some combinations of base classifiers in the MtCE exceed the performance of all of the base classifiers (see the bold-green scores in Table 4). For example, accuracy of the MtCE(LR-MLP) for Ott et al.'s = 88%; in comparison, in Table 5, accuracy of LR and MLP are 87.13% and 87.56%. However, not all combinations improved and supported each other as we hoped. A few weak classifiers degraded the performance of stronger classifiers when combined with these. This gave rise to the result that the performance of the MtCE was in the middle of the performance of its base classifiers; for example, LSVM accuracy for Ott et al.'s = 85.88% while MLP = 87.56%, and the combination of both in MtCE(LSVM-MLP) = 87.50%. This implies that the LSVM degraded the performance of MLP, and given this, rather than implementing our proposed ensemble, it would be better to use the MLP directly. However, in the same case, MtCE(LSVM-MLP) for Ott et al.'s improved the recall, from 86.21% (LSVM) and 88.37% (MLP) to 90.02%. In a case such as this, using our ensemble would be acceptable since the recall in binary classification/detection is the same as the accuracy of the positive class or the main class (in our problem, the fake review class). Thus, high recall means the ability of the ensemble to detect positive classes is also high.

Another finding from this set of experiments is that different datasets might need different base classifier combinations. For example, the MtCE(MLP-NB) that performed best for Ott et al.'s dataset performed worst for Li et al.'s doctor dataset (increasing the recall but decreasing all other measurements). The best combinations for Li et al.'s doctor dataset was MtCE(LSVM-MLP); for Li et al.'s hotel dataset was MtCE(LR-MLP-DT-NB); and for Li et al.'s restaurant dataset was MtCE(LR-LSVM-NB). A closer inspection of Table 5 shows that the base classifiers of MtCE's best combination for a particular dataset mostly performed well on the same dataset too, when they were used in stand-alone modes. For example, on Ott's MtCE(MLP-NB), the MLP and NB also performed well on Ott's dataset, although they were not as good as when combined in the MTCE. This observation is valid on other combinations for other datasets, except the DT on Li et al.'s hotel dataset, which was not performing well by itself but performed best in combination with other base classifiers in the MtCE. This indicates that while combining good classifiers could produce better results, infusing a weak classifier sometimes could also boost the results of MtCE.

We found that the MtCE performed better than other ensembles of homogenous classifiers, as indicated by comparing the results in Table 5 (numbers 7–10) to the MtCE results in Table 4. These facts suggest that, when performed correctly, the combination of multi-type classifiers could cover the weakness of each and outperform the combination of a homogenous type of ensemble classifiers such as RF, BP, AB or GB. However, for the cases of RF, BP and AB, we suspect the cause is also because these ensembles use DT as their base classifiers/predictors. DT's performance was not good on all datasets, which therefore affected the performance of its ensemble variants.

While deep learning CNN as a stand-alone classifier did not perform as well as other single classifiers, somewhat surprisingly, it performed exceptionally well when combined with other classifiers in the MtCE. This increase in accuracy was significant; for example, MtCE(CNN-LR-LSVM-MLP-NB) is 7.93% in Li et al.'s hotel dataset to 18.45% in Li et al.'s doctor dataset. The improvement in recall was also good, with a minimum of 4.78% in Li et al.'s doctor dataset to 12.97% in Li et al.'s restaurant dataset. Therefore, the deep learning CNN could be combined perfectly with other ML single classifiers as base classifiers for the MtCE; it gave good results and improved all other base classifiers' performances as well. This is another indication that combining different classifiers with their strength and weakness in MtCE would improve the overall performance. It is because the weakness of one base classifier could be covered by the strength of the other base classifiers.

5.2 EFFECTS OF MTCE PARAMETERS

In this section, we discuss several input parameters of the MtCE that may affect the performance of this proposed model. All of the settings are the same as in Section 5.1 except the parameter we are testing. For these experiments, we chose four base classifier combinations of the MtCE in Table 4—that is, MtCE(LSVM-MLP), MtCE(LSVM-NB), MtCE(LR-LSVM-NB), MtCE(LR-MLP-DT-NB) and MtCE(LR-LSVM-MLP-DT-NB)—as these are the five combinations of base classifier types that have performed the best across all four datasets. We ran all experiments for the parameter testing using Ott et al.'s dataset only, since these five MtCE combination settings perform well in this dataset (see the bold-green scores in Table 4). However, after finding the ideal set of parameters, we ran them with the other datasets and compared the results with our previous approach implemented on the same datasets.

5.2.1 TOTAL NUMBER OF BASE CLASSIFIERS

First, we investigated the effect of base classifiers' total numbers. Here, we ran experiments on all five combinations with the total number of base classifiers varying from 5 to 100. Results can be seen in Figure 4 for accuracy, Figure 5 for precision, Figure 6 for recall and Figure 7 for F-measure. From these figures, we can see that accuracy, precision, recall and F-measure were increasing at first, then, after 20 classifiers, all scores fluctuated around a number ± 1.5%. Accuracy fluctuated at 88.5%, precision at 88% and recall at 90%, except for MtCE(LSVM-MLP), which was 1% lower. Intuitively, these fluctuations after 20 classifiers mean that adding more classifiers would not improve the measurements a lot. However, since the case is only for fake review detection on Ott et al.'s dataset, we will let the user decide on the number of base classifiers needed for their problem. Next, since we think a 1.5% difference is quite significant, we used the best accuracy results for each MtCE combination for the next batch of experiments. These combinations are 20 classifiers for MtCE(LSVM-MLP), 55 classifiers for MtCE(LR-LSVM-NB), 80 classifiers for MtCE(LR-MLP-DT-NB) and 85 classifiers for MtCE(LR-LSVM-MLP-DT-NB).

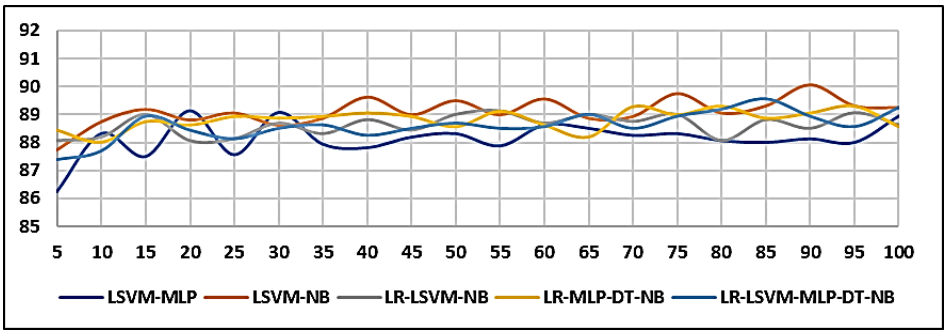

Figure 4: The accuracy (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis)

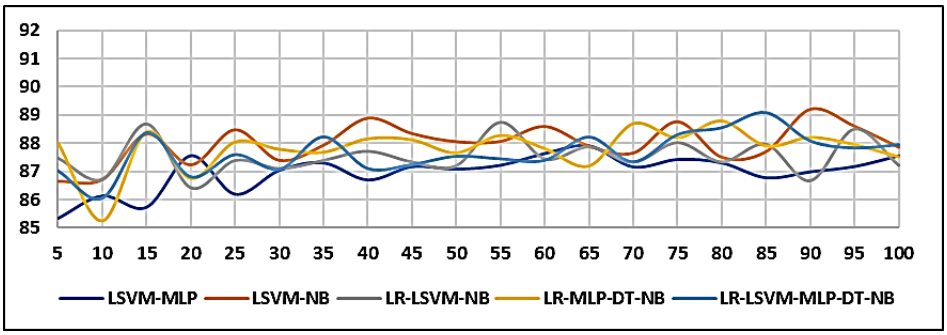

Figure 5: The precision (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis)

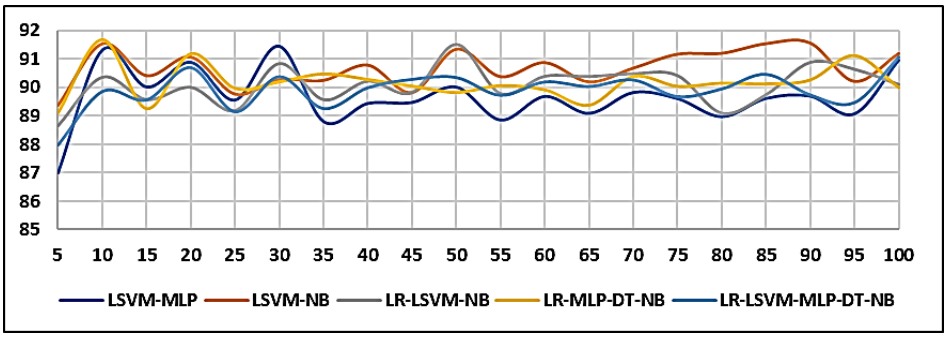

Figure 6: The recall (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis)

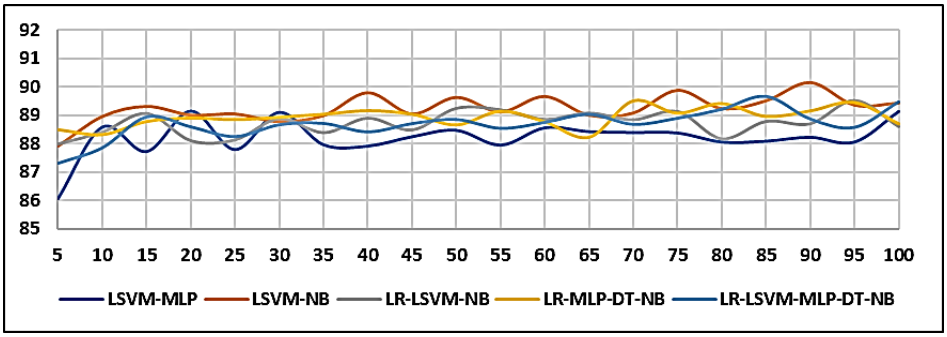

Figure 7: The F-measure (y-axis, in %) of MtCE with 5–100 total base classifiers (x-axis)

5.2.2 EQUALLY SPLIT OR RANDOMLY ASSIGNED BASE CLASSIFIER TYPE

To investigate the equally split vs randomly assigned base classifier type, we ran the experiments 10 times for each equally split and randomly assigned setting for each MtCE combination. The average results and their standard deviation can be seen in Table 6. This table shows that the equally split option gave slightly better average results for all measurements relative to the random option. This implies that when the total of each base classifier is equal, they support each other, and the strength of one type of classifier covers the weakness of the other type when combined in the MtCE. On the other hand, the random option could make one or two base classifiers more numerous than the others, so that they dominated the detection process. The standard deviation of all measurement scores below 1 means the variation of the results from 10 runs was low and close to each other. In more detail, we can see that the standard deviation of 2-combination, the MtCE(LSVM-MLP), on the equally split setting was lower than the random assignment. This is expected since, in an equally split setting, each base classifier total is the same for all 10 runs, while in the random option, this varies. However, the exact opposite happened in 5-combination, the MtCE(LR-LSVM-MLP-DT-NB). Here, the standard deviation of the equally split setting was higher than the random option, implying that the randomness might make the base classifiers more homogenous than split equally between its 5 base classifier types. For the 3- and 4-combinations, the standard deviation scores are similar.

Table 6. Results of equally split vs randomly assigned type of MtCE base classifiers on 10 runs of the 10-fold CV process.

Base Classifier Combination	Equally Split or Randomly Assigned	Accuracy		Precision		Recall		F-measure	
		Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev
LSVM-MLP	Equal	88.21	0.38	86.76	0.35	90.23	0.56	88.39	0.41
	Random	87.82	0.60	86.41	0.59	89.81	0.69	88.00	0.60
LSVM- NB	Equal	89.44	0.36	88.20	0.50	91.12	0.39	89.57	0.37
	Random	89.37	0.19	88.16	0.38	90.93	0.35	89.47	0.22

LR-LSVM-NB	Equal	88.97	0.35	88.14	0.48	90.06	0.42	89.04	0.35
	Random	88.72	0.37	87.79	0.41	89.89	0.49	88.78	0.38
LR-MLP-DT-NB	Equal	89.18	0.32	88.34	0.44	90.24	0.39	89.21	0.29
	Random	89.03	0.35	88.19	0.46	90.12	0.38	89.09	0.35
LR-LSVM-MLP-DT-NB	Equal	88.28	0.66	87.37	0.64	89.59	0.86	88.41	0.67
	Random	88.12	0.37	87.10	0.39	89.49	0.49	88.21	0.39

5.2.3 BOOTSTRAP EFFECT

The subsequent experiments aimed to investigate the bootstrap parameter. We used fixed parameters as above, except the bootstrap setting ranged from 0.25 to 1 (no bootstrapping). As can be seen in Figure 8, the best accuracy, precision and recall were often achieved when the bootstrap setting was 0.5, where every base classifier was trained using 50% of training samples randomly. This indicates that bootstrap setting 0.5 is ideal for the tested MtCE settings, i.e., the five best base-classifier combinations. Therefore, intuitively, it can be used as the default bootstrap setting of MtCE. Based on these findings, for the next set of experiments, we set the bootstrap to be 0.5.

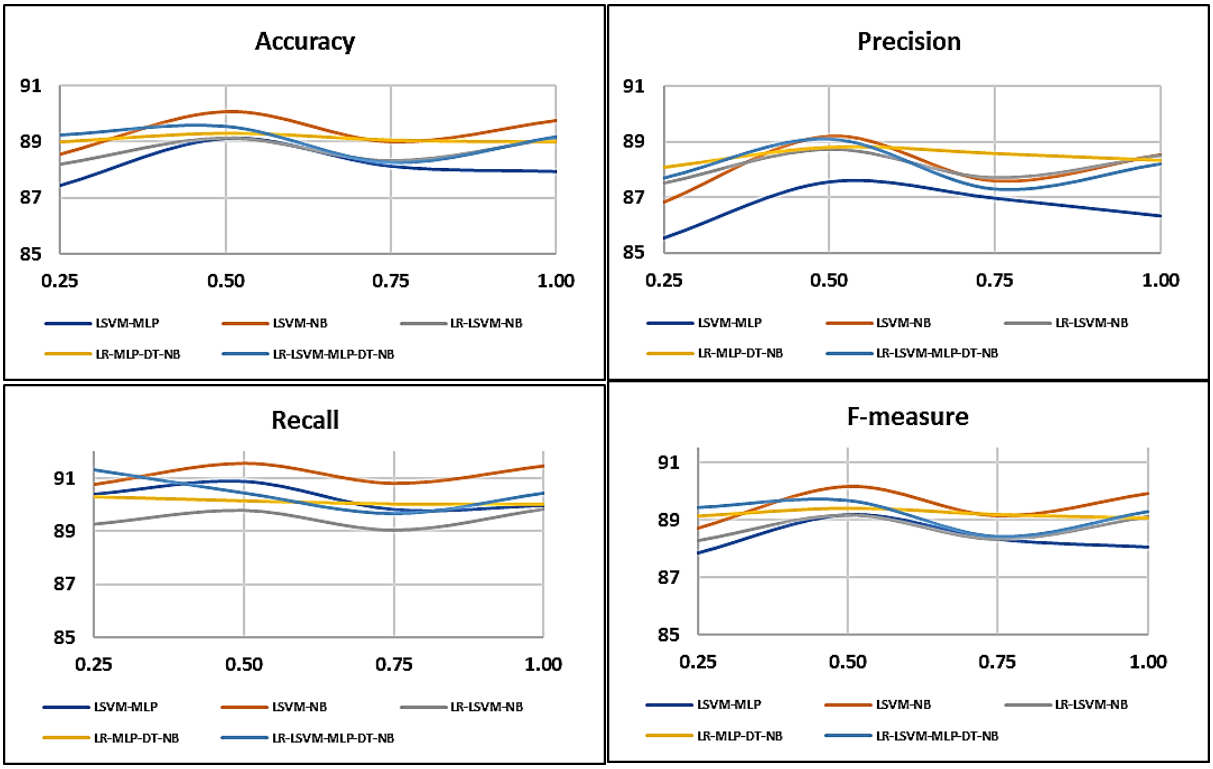

Figure 8: The accuracy, precision, recall and F-measure (y-axis, in %) of the MtCE with bootstrap settings from 0.25 to 1.00 (x-axis)

5.2.4 FINAL OUTPUT: MAJORITY VOTE OR CLASS PRIORITY THRESHOLD

The last parameter to test is how the output of base classifiers should be processed. There are two options in terms of transferring the output of base classifiers to the final output: majority vote and priority class. To test the effect of priority, we set the priority to the positive class (fake class), in the range between absolute priority and 45% priority. As mentioned above, absolute priority means the final result will be recorded as positive/fake if one or more of the base classifiers records this. Percentage priority means the final result will be recorded as positive/fake if the number of base classifiers with this result is the same or more than the percentage threshold. The results of the experiments can be seen in Figure 9.

As per Figure 9, the priority setting increases the recall but decreases other measurements. Moreover, as noted, the higher the recall, the more accurate the positive class (fake class) detection. Thus, we can conclude that the priority setting is helpful when samples of positive class are scarce or when the focus of research is on anomaly detection. The key is how high we choose the percentage of priority so that the increase in recall does not greatly decrease the other measurements. For example, in Figure 9, for 25% priority of MtCE(LSVM-MLP), where recall increased by 5.05%, the accuracy, precision and F-measure decreased by 2.75%, 7.15% and 1.70%, respectively.

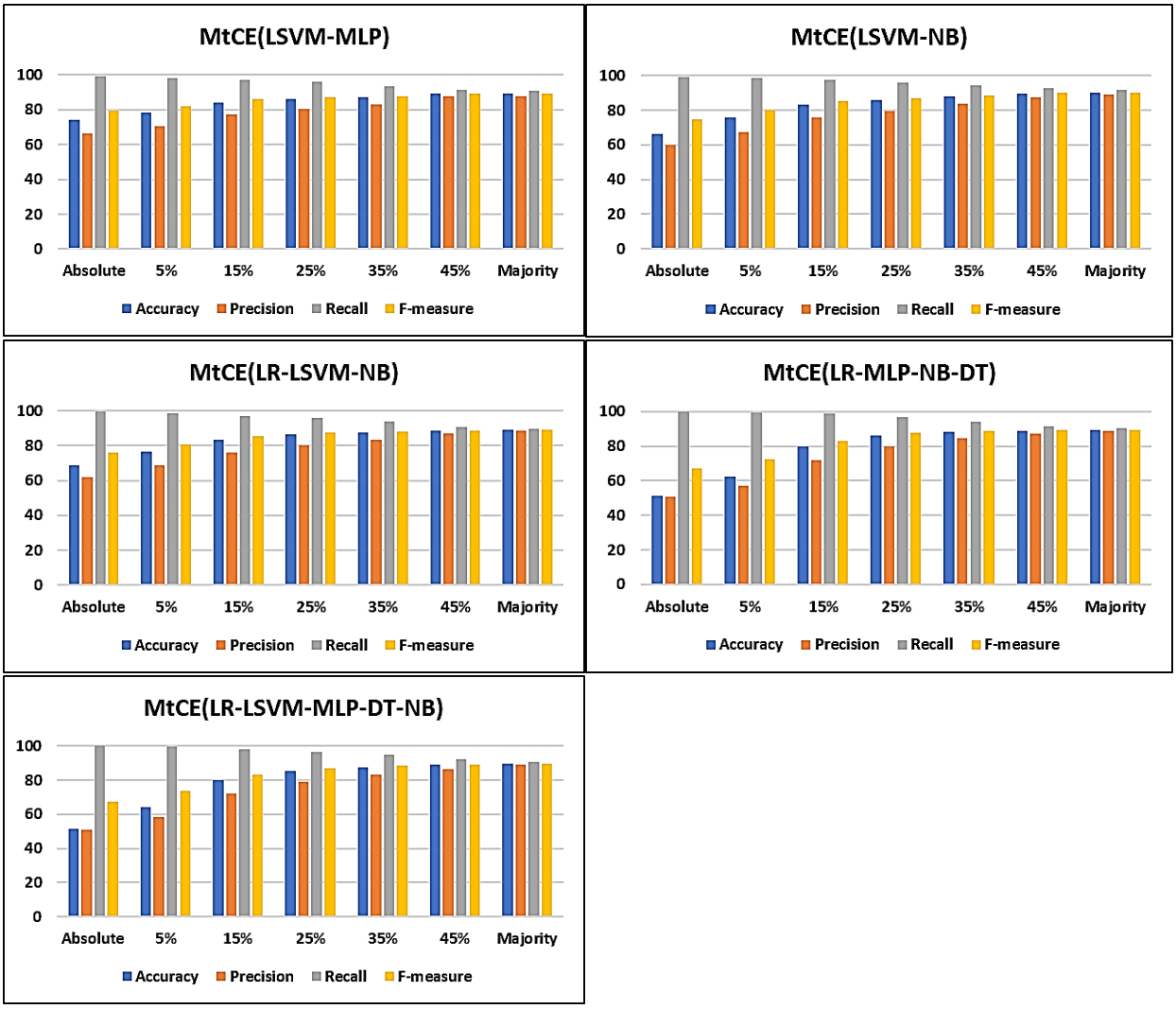

Figure 9: Results of final target experiments, in terms of accuracy, precision, recall and F-measure (y-axis, in %), from absolute priority to 45% priority, and majority (x-axis)

5.3 COMPARISON TO OTHER STUDIES ON THE SAME DATASETS

As commonly presented in similar studies [42; 55; 57; 72], in this section, we present the MtCE's and other models' best results in light of various considerations. Many factors can influence results, such as the measurement formulas or tools, datasets used, number of records involved in experiments, and how the experiments were conducted (e.g., CV vs traditional train-test-validation split). Therefore, the results presented in Table 7 should be read with the following considerations in mind:

1. We present the MtCE's and other studies' results based on the same datasets (see Table 2). We used all samples that could be processed from each dataset. Note that some studies used only a subset of the dataset

or split the dataset into subsets and measured each subset differently. All of our experiments were conducted using 10-fold CV.

2. It is impossible to ensure that all the studies used the same formulas to measure accuracy, precision, recall and F-measure. Therefore, MtCE results were measured using widely used formulas for binary target prediction (see Table 3). All measurements were in the 'macro' average setting using scikit-learn measurement components [51].
3. The original results from the dataset creators were used as the baseline (see the underlined description and scores in Table 7. We present only the best result for each dataset from each study.

Table 7: Best performance of the MtCE and other studies for the same datasets

No	Dataset	Description	Acc (%)	Pre (%)	Rec (%)	F1 (%)
1	Hotel (various) [49]	Ott et al. [49], SVM on positive sentiment subset	88.4	89.1	87.5	88.3
		Fusilier et al. [27], PU-L modified	-	85.2	72.8	78
		Rout et al. [57], DT	92.11	-	-	-
		Etaiwi and Naymat [21], SVM, all preproc. steps	85	51	86	-
		Zhang et al. [72], DRI-RCNN, 0.9 T.P.	-	-	-	86.59
		Budhi et.al. [12], GB, over-sampling	70.62	67.58	81.29	73.8
		*MtCE(LSVM-NB, 90, Equal, 0.5, Majority)**	*90.06*	***89.19***	***91.56***	***90.16***
2	Hotel (various) [41]	Li et al. [41], OvR SAGE-Unigram	81.8	81.2	84	-
		Ren and Ji [55], Integrated, Employee/Turker subset	92.6	-	-	90.1
		Li et al. [42], SWNN	-	84.1	87	85
		Budhi et.al. [12], GB, under-sampling	71.29	73.61	82.79	77.93
		*MtCE(LR-MLP-DT-NB, 15, Equal, 0.5, Majority)**	*87.82*	***86.67***	***93.26***	*89.81*
3	Restaurant (various) [41]	Li et al. [41], OvR SAGE-Unigram	81.7	84.2	81.6	-
		Ren and Ji [55], Integrated, Customer/Turker subset	86.9	-	-	86.8
		Li et al. [42], SWNN		83.3	88.2	81
		Budhi et.al. [12], GB, over-sampling	72.44	71.23	75.16	73.14
		*MtCE(LR-MLP-DT-NB , 80, Equal, 0.5, Majority)**	*86.07*	***84.47***	***88.89***	*86.37*
4	Doctor (various) [41]	Li et al. [41], OvR SAGE-Unigram	74.5	77.2	70.1	-
		Ren and Ji [55], Integrated, Customer/Turker subset	76	-	-	74.1
		Li et al. [42], SWNN	-	83.7	87.6	82.9
		Budhi et.al. [12], GB, over-sampling	73.35	79.44	86.94	83.02
		*MtCE(LSVM-MLP, 15, Equal, 0.5, Majority)**	***86.56***	***87.05***	***93.12***	***89.88***

* MtCE(<type of base classifiers>, <total base classifiers>, <equally split or random assign of base classifiers>, <bootstrap percentage>, <majority or priority output>)

We do not claim that the MtCE is better or worse than methods in other studies since several factors can affect results. However, we can confidently compare the MtCE to our previous study in fake review detection [12] because the measurement formulas are similar. As is evident in Table 7, compared with our previous approach, the MtCE is superior. Accuracy improves from a minimum of 13.2% in Li et al.'s doctor dataset to a maximum of 19.4% in Ott et al.'s dataset; precision improves from 7.6% to 21.6% and recall from 6.2% to 13.7%.

6 CONCLUSION AND FUTURE WORK

From the experiments above, we can conclude that our proposed ensemble, the MtCE, was able to adequately detect fake or fraudulent reviews, which have become an acute problem on e-commerce platforms. Compared with our previous approach, the MtCE improved accuracy up to 19.4%, precision up to 21.6% and recall up to 13.7%.

Not all combinations of base classifier types produced the expected result of an improvement in the performance on all base classifiers of the MtCE. In some combinations, weak classifiers dragged down the performance of stronger classifiers, especially in terms of accuracy measurements. However, when the combinations were correct, the MtCE

produced better results than its base classifiers in stand-alone modes. It is important to note that while accuracy was not on top, recall was the highest in many cases. This is a positive result since, in binary classification, recall is the same as accuracy of a positive case. In this case, the MtCE was able to better detect the fake review class. Comparison with well-known ensembles, such as the RF, AB, BP and GB, showed that some correct combination of the MtCE could outperform the other ensembles, proving that combining multi-type classifiers in one ensemble process performed better than combining the same classifiers, as is usually done in other ensemble methods.

We proved that DL methods such as the CNN model could be combined with ML counterparts as base classifiers to give better results than if run alone. The number of base classifiers affected performance detection; however, accuracy and other measurements oscillated after 20 base classifiers. While reducing the amount of data for the training process, the bootstrap setting could also affect the performance of the MtCE. From the experiments, we found that 50% bootstrapping gives better results than other percentages, including the non-bootstrap (bootstrap setting 100%). While the majority vote setting for the output offers a fair chance for each class to be detected, the priority threshold boosts detection of the prioritised class. This would be useful if the dataset is imbalanced or the MtCE is used to detect/classify anomalies.

Despite the extensive experimental studies presented in this paper, there remain opportunities/possibilities for further improvement. First, we are aware that multiple types of base classifiers will increase the complexity of the ensemble and increase the cost and time processing for parameter tuning. To overcome this problem, in the near future, we plan to expand on our previously proposed algorithms for ensemble parameter optimisation, namely the Multi-level Particle Swarm Optimisation (PSO) and its parallel version [8]. These nature-inspired algorithms are based on a low-cost PSO algorithm. They were designed to optimise the parameters of an ensemble model, such as BP and AB, choose its base classifier type, and optimise the parameters of these base classifier candidates. Other avenues for future work include investigating the implementation of the MtCE for multi-class problems, heavily imbalanced datasets, and anomaly detection.

REFERENCES

[1] AKRAM, A.U., KHAN, H.U., IQBAL, S., IQBAL, T., MUNIR, E.U., and SHAFI, M., 2018. Finding rotten eggs: A review spam detection model using diverse feature sets. *KSII Transactions on Internet and Information Systems 12*, 10. DOI= http://dx.doi.org/10.3837/tiis.2018.10.026.

[2] ALSUBARI, S.N., SHELKE, M.B., and DESHMUKH, S.N., 2020. Fake reviews identification based on deep computational linguistic features. *International Journal of Advanced Science and Technology 29*, 8s, 3846-3856.

[3] BARBADO, R., ARAQUE, O., and IGLESIAS, C.A., 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management 56*, 4, 1234-1244. DOI= http://dx.doi.org/10.1016/j.ipm.2019.03.002.

[4] BING, L., WONG, T.-L., and LAM, W., 2016. Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. *ACM Transactions on Internet Technology 16*, 2, 1-17. DOI= http://dx.doi.org/10.1145/2857054.

[5] BREIMAN, L., 1996. Bagging predictors. *Machine Learning 24*, 2, 123-140. DOI= http://dx.doi.org/https://doi.org/10.1007/bf00058655.

[6] BREIMAN, L., 1999. Pasting Small Votes for Classification in Large Databases and On-Line. *Machine Learning 36*, 1 (1999/07/01), 85-103. DOI= http://dx.doi.org/10.1023/A:1007563306331.

[7] BREIMAN, L., 2001. Random forests. *Machine Learning 45*, 1, 5-32. DOI= http://dx.doi.org/https://doi.org/10.1023/a:1010933404324.

[8] BUDHI, G.S., CHIONG, R., and DHAKAL, S., 2020. Multi-level particle swarm optimisation and its parallel version for parameter optimisation of ensemble models: a case of sentiment polarity prediction. *Cluster Computing 23*, 3371-3386. DOI= http://dx.doi.org/https://doi.org/10.1007/s10586-020-03093-3.

[9] BUDHI, G.S., CHIONG, R., PRANATA, I., and HU, Z., 2017. Predicting rating polarity through automatic classification of review texts. In *Proceedings of the 2017 IEEE Conference on Big Data and Analytics (ICBDA)*, Kuching, Malaysia, 19-24. DOI= http://dx.doi.org/https://doi.org/10.1109/ICBDAA.2017.8284101.

[10] BUDHI, G.S., CHIONG, R., PRANATA, I., and HU, Z., 2021. Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis. *Archives of Computational Methods in Engineering 28*, 2543–2566. DOI= http://dx.doi.org/https://doi.org/10.1007/s11831-020-09464-8.

[11] BUDHI, G.S., CHIONG, R., and WANG, Z., 2021. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications 80*, 13079-13097. DOI= http://dx.doi.org/https://doi.org/10.1007/s11042-020-10299-5.

[12] BUDHI, G.S., CHIONG, R., WANG, Z., and DHAKAL, S., 2021. Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. *Electronic Commerce Research and Applications 47, 101048*. DOI= http://dx.doi.org/https://doi.org/10.1016/j.elerap.2021.101048.

[13] CAMPBELL, C. and YING, Y., 2011. *Learning with Support Vector Machines*. Morgan & Claypool.

[14] CARDOSO, E.F., SILVA, R.M., and ALMEIDA, T.A., 2018. Towards automatic filtering of fake reviews. *Neurocomputing 309*, 106-116. DOI= http://dx.doi.org/10.1016/j.neucom.2018.04.074.

[15] CHANG, C.C. and LIN, C.J., 2011. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3, 1-27. DOI= http://dx.doi.org/https://doi.org/10.1145/1961189.1961199.

[16] CHIONG, R., BUDHI, G.S., and DHAKAL, S., 2021. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems 36*, 6. DOI= http://dx.doi.org/https://doi.org/10.1109/MIS.2021.3093660.

[17] CHIONG, R., BUDHI, G.S., DHAKAL, S., and CHIONG, F., 2021. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine 135, 104499*. DOI= http://dx.doi.org/https://doi.org/10.1016/j.compbiomed.2021.104499.

[18] DENG, X., LI, Y., WENG, J., and ZHANG, J., 2019. Feature selection for text classification: A review. *Multimedia Tools and Applications 78*, 3, 3797-3816. DOI= http://dx.doi.org/https://doi.org/10.1007/s11042-018-6083-5.

[19] DONG, M., YAO, L., WANG, X., BENATALLAH, B., HUANG, C., and NING, X., 2018. Opinion fraud detection via neural autoencoder decision forest. *Pattern Recognition Letters*. DOI= http://dx.doi.org/10.1016/j.patrec.2018.07.013.

[20] ELMOGY, A.M., TARIQ, U., and MOHAMMED, A.I.A., 2021. Fake Reviews Detection using Supervised Machine Learning. *International Journal of Advanced Computer Science and Applications 12*, 1, 601-. DOI= http://dx.doi.org/https://dx.doi.org/10.14569/IJACSA.2021.0120169.

[21] ETAIWI, W. and NAYMAT, G., 2017. The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science 113*, 273-279.

[22] FELBERMAYR, A. and NANOPOULOS, A., 2016. The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing 36*, 60-76. DOI= http://dx.doi.org/10.1016/j.intmar.2016.05.004.

[23] FENG, V.W. and HIRST, G., 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of International Joint Conference on Natural Language Processing*, Nagoya, Japan, 338-346.

[24] FREUND, Y. and SCHAPIRE, R.E., 1997. A Decision-Theoretic Generalisation of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences 55*, 1 (1997/08/01/), 119-139. DOI= http://dx.doi.org/https://doi.org/10.1006/jcss.1997.1504.

[25] FRIEDMAN, J.H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 5, 1189-1232.

[26] HAZIM, M., ANUAR, N.B., AB RAZAK, M.F., and ABDULLAH, N.A., 2018. Detecting opinion spams through supervised boosting approach. *PLoS ONE 13*, 6, e0198884. DOI= http://dx.doi.org/10.1371/journal.pone.0198884.

[27] HERNÁNDEZ FUSILIER, D., MONTES-Y-GÓMEZ, M., ROSSO, P., and GUZMÁN CABRERA, R., 2015. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management 51*, 4, 433-443. DOI= http://dx.doi.org/10.1016/j.ipm.2014.11.001.

[28] HEYDARI, A., TAVAKOLI, M., and SALIM, N., 2016. Detection of fake opinions using time series. *Expert Systems with Applications 58*, 83-92. DOI= http://dx.doi.org/10.1016/j.eswa.2016.03.020.

[29] HEYDARI, A., TAVAKOLI, M.A., SALIM, N., and HEYDARI, Z., 2015. Detection of review spam: A survey. *Expert Systems with Applications 42*, 7, 3634-3642. DOI= http://dx.doi.org/10.1016/j.eswa.2014.12.029.

[30] HU, Z., CHIONG, R., PRANATA, I., BAO, Y., and LIN, Y., 2019. Malicious web domain identification using online credibility and performance data by considering the class imbalance issue. *Industrial Management & Data Systems 119*, 3, 676-696. DOI= http://dx.doi.org/https://doi.org/10.1108/IMDS-02-2018-0072.

[31] HU, Z., CHIONG, R., PRANATA, I., SUSILO, W., and BAO, Y., 2016. Identifying malicious web domains using machine learning techniques with online credibility and performance data. In *Proceedings of Congress on Evolutionary Computation (CEC)*, Vancouver, Canada, 5186-5194.

[32] JACOBS, R.A., JORDAN, M.I., NOWLAN, S.J., and HINTON, G.E., 1991. Adaptive Mixtures of Local Experts. *Neural Computation 3*, 1, 79-87. DOI= http://dx.doi.org/https://doi.org/10.1162/neco.1991.3.1.79.

[33] JAVED, M.S., MAJEED, H., MUJTABA, H., and BEG, M.O., 2021. Fake reviews classification using deep learning ensemble of shallow convolutions. *Journal of Computational Social Science 4*, 2, 883-902. DOI= http://dx.doi.org/10.1007/s42001-021-00114-y.

[34] JIANG, Z., SAINJU, A.M., LI, Y., SHEKHAR, S., and KNIGHT, J., 2019. Spatial ensemble learning for heterogeneous geographic data with class ambiguity. *ACM Transactions on Intelligent Systems and Technology 10*, 4, 1-25. DOI= http://dx.doi.org/10.1145/3337798.

[35] JOHNSON, R.W., 2001. An introduction to the Bootstrap. *Teaching Statistics 23*, 2, 49-54. DOI= http://dx.doi.org/https://doi.org/10.1111/1467-9639.00050.

[36] KERAS, 2019. Keras: The Python Deep Learning library.

[37] KONTSCHIEDER, P., FITERAU, M., CRIMINISI, A., and BUL`O, S.R., 2015. Deep Neural Decision Forests. In *2015 IEEE International Conference on Computer Vision (ICCV)* IEEE, Santiago, Chile, 1467-1475.

[38] KUMAR, A., GOPAL, R.D., SHANKAR, R., and TAN, K.H., 2022. Fraudulent review detection model focusing on emotional expressions and explicit aspects: investigating the potential of feature engineering. *Decision Support Systems 155*. DOI= http://dx.doi.org/10.1016/j.dss.2021.113728.

[39] KUMAR, N., VENUGOPAL, D., QIU, L., and KUMAR, S., 2018. Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems 35*, 1, 350-380. DOI= http://dx.doi.org/10.1080/07421222.2018.1440758.

[40] LECUN, Y., BOTTOU, L., BENGIO, Y., and HAFFNER, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11, 2278-2324. DOI= http://dx.doi.org/https://doi.org/10.1109/5.726791.

[41] LI, J., OTT, M., CARDIE, C., and HOVY, E., 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, 1566-1576.

[42] LI, L., QIN, B., REN, W., and LIU, T., 2017. Document representation and feature combination for deceptive spam review detection. *Neurocomputing 254*, 33-41. DOI= http://dx.doi.org/10.1016/j.neucom.2016.10.080.

[43] MALBON, J., 2013. Taking fake online consumer reviews seriously. *Journal of Consumer Policy 36*, 2, 139-157. DOI= http://dx.doi.org/10.1007/s10603-012-9216-7.

[44] MANNING, C.D., RAGHAVAN, P., and SCHUETZE, H., 2008. Naïve Bayes text classification. In *Introduction to Information Retrieval* Cambridge University Press, 234-265.

[45] MARTENS, D. and MAALEJ, W., 2019. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering 24*, 6, 3316-3355. DOI= http://dx.doi.org/10.1007/s10664-019-09706-9.

[46] MENARD, S., 2010. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE, Los Angeles.

[47] MOHAMMADI, Y., SALARPOUR, A., and CHOUHY LEBORGNE, R., 2021. Comprehensive strategy for classification of voltage sags source location using optimal feature selection applied to support vector machine and ensemble techniques. *International Journal of Electrical Power & Energy Systems 124*. DOI= http://dx.doi.org/10.1016/j.ijepes.2020.106363.

[48] NORVIG, P., 2016. How to write a spelling corrector.

[49] OTT, M., CARDIE, C., and HANCOCK, J.T., 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, US, 497-501.

[50] OTT, M., CHOI, Y., CARDIE, C., and HANCOCK, J.T., 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings ofthe 49th Annual Meeting ofthe Association for Computational Linguistics*, Portland, Oregon, 309-319.

[51] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., and DUCHESNAY, É., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 85 (May 08), 2825-2830.

[52] POLIKAR, R., 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine 6*, 3, 21-45. DOI= http://dx.doi.org/10.1109/MCAS.2006.1688199.

[53] QUINLAN, J.R., 1986. Induction of decision trees. *Machine Learning 1*, 1 (March 01), 81-106. DOI= http://dx.doi.org/https://doi.org/10.1007/bf00116251.

[54] REN, Q., LI, M., and HAN, S., 2019. Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. *Big Earth Data 3*, 1, 8-25. DOI= http://dx.doi.org/https://doi.org/10.1080/20964471.2019.1572452.

[55] REN, Y. and JI, D., 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences 385-386*, 213-224. DOI= http://dx.doi.org/10.1016/j.ins.2017.01.015.

[56] REN, Y., ZHANG, L., and SUGANTHAN, P.N., 2016. Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]. *IEEE Computational Intelligence Magazine 11*, 1, 41-53. DOI= http://dx.doi.org/10.1109/mci.2015.2471235.

[57] ROUT, J.K., SINGH, S., JENA, S.K., and BAKSHI, S., 2016. Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications 76*, 3, 3187-3211. DOI= http://dx.doi.org/10.1007/s11042-016-3819-y.

[58] RUMELHART, D.E., HINTON, G.E., and WILLIAMS, R.J., 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* MIT Press, 318-362.

[59] SAVAGE, D., ZHANG, X., YU, X., CHOU, P., and WANG, Q., 2015. Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications 42*, 22, 8650-8657. DOI= http://dx.doi.org/10.1016/j.eswa.2015.07.019.

[60] SCHAPIRE, R.E., 1990. The strength of weak learnability. *Machine Learning 5*, 2 (1990/06/01), 197-227. DOI= http://dx.doi.org/10.1007/BF00116037.

[61] SHAN, G., ZHOU, L., and ZHANG, D., 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems 144*. DOI= http://dx.doi.org/10.1016/j.dss.2021.113513.

[62] SHIN, J., YOON, S., KIM, Y., KIM, T., GO, B., and CHA, Y., 2021. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics 61*. DOI= http://dx.doi.org/10.1016/j.ecoinf.2020.101202.

[63] SUN, C., DU, Q., and TIAN, G., 2016. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering 2016*, 1-7. DOI= http://dx.doi.org/10.1155/2016/4935792.

[64] TANG, X., QIAN, T., and YOU, Z., 2020. Generating behavior features for cold-start spam review detection with adversarial learning. *Information Sciences 526*, 274-288. DOI= http://dx.doi.org/10.1016/j.ins.2020.03.063.

[65] WAHYUNI, E.D. and DJUNAIDY, A., 2016. Fake review detection from a product review using modified method of iterative computation framework. In *Proceedings of MATEC Web of Conferences 58, 03003*. DOI= http://dx.doi.org/10.1051/matec.

[66] WANG, X., HE, K.L.S., and ZHAO, J., 2016. Learning to Represent Review with Tensor Decomposition for Spam Detection. In *Proceedings ofthe 2016 Conference on Empirical Methods in Natural Language Processing*, Texas, US, 866-875.

[67] WOLPERT, D.H., 1992. Stacked generalisation. *Neural Networks 5*, 2 (1992/01/01/), 241-259. DOI= http://dx.doi.org/https://doi.org/10.1016/S0893-6080(05)80023-1.

[68] WU, Y., NGAI, E.W.T., WU, P., and WU, C., 2020. Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems 132*. DOI= http://dx.doi.org/10.1016/j.dss.2020.113280.

[69] XIE, F., FAN, H., LI, Y., JIANG, Z., MENG, R., and BOVIK, A., 2017. Melanoma Classification on Dermoscopy Images Using a Neural Network Ensemble Model. *IEEE Trans Med Imaging 36*, 3 (Mar), 849-858. DOI= http://dx.doi.org/10.1109/TMI.2016.2633551.

[70] XIONG, T., BAO, Y., HU, Z., and CHIONG, R., 2015. Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. *Information Sciences 305*, 77-92. DOI= http://dx.doi.org/https://doi.org/10.1016/j.ins.2015.01.029.

[71] ZHANG, D., ZHOU, L., KEHOE, J.L., and KILIC, I.Y., 2016. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems 33*, 2, 456-481. DOI= http://dx.doi.org/https://doi.org/10.1080/07421222.2016.1205907.

[72] ZHANG, W., DU, Y., YOSHIDA, T., and WANG, Q., 2018. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management 54*, 4, 576-592. DOI= http://dx.doi.org/10.1016/j.ipm.2018.03.007.

7. Paper accepted news from Correspondent Author (23 October 2022)

Dr. Raymond Chiong

toit 3 of 13

Really 6:19 PM

Yes. 6:21 PM

10/23/2022

Dear Raymond Chiong,

Your publication has reached our "Just Accepted" state and the Author's Accepted Manuscript (AAM) is now available in the ACM Digital Library, https://doi.acm.org?doi=3568676 .

Within approximately 2-4 weeks you will receive a separate link from the Aptara journals support team to review and make any changes (if needed) to the AAM proof of your article at that time.

A multi-type classifier ensemble for detecting fake reviews through textual-based feature extraction
Gregorius Satia Budhi, Raymond Chiong

ACM Transactions on Internet Technology 4:24 AM

👍

A Multi-type Classifier Ensemble for Detecting Fake Reviews Through Textual-based Feature Extraction

GREGORIUS SATIA BUDHI, School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW, Australia, and Informatics Department, Petra Christian University, Surabaya, Indonesia
RAYMOND CHIONG, School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW, Australia

The financial impact of online reviews has prompted some fraudulent sellers to generate fake consumer reviews for either promoting their products or discrediting competing products. In this study, we propose a novel ensemble model—the **Multi-type Classifier Ensemble (MtCE)**—combined with a textual-based featuring method, which is relatively independent of the system, to detect fake online consumer reviews. Unlike other ensemble models that utilise only the same type of single classifier, our proposed ensemble utilises several customised machine learning classifiers (including deep learning models) as its base classifiers. The results of our experiments show that the MtCE can adequately detect fake reviews, and that it outperforms other single and ensemble methods in terms of accuracy and other measurements for all the relevant public datasets used in this study. Moreover, if set correctly, the parameters of MtCE, such as base-classifier types, the total number of base classifiers, bootstrap, and the method to vote on output (e.g., majority or priority), can further improve the performance of the proposed ensemble.

CCS Concepts: • **Computing methodologies** → **Machine learning**;

Additional Key Words and Phrases: Fake review detection, online commerce security, novel ensemble model, machine learning, deep learning

ACM Reference format:
Gregorius Budhi and Raymond Chiong. 2023. A Multi-type Classifier Ensemble for Detecting Fake Reviews Through Textual-based Feature Extraction. *ACM Trans. Internet Technol.* 23, 1, Article 16 (April 2023), 24 pages.
https://doi.org/10.1145/3568676

1 INTRODUCTION

Fake review detection is a hot research topic in natural language processing [55]. A fake review is a consumer review of a product with fictitious opinions written deliberately to sound authentic, for commercial motives; i.e., to promote or damage the reputation of the reviewed product [42, 50]. There is a significant possibility that fake reviews distort the actual evaluation of a product

Authors' addresses: G. S. Budhi, School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia, and Informatics Department, Petra Christian University, Surabaya 60236, Indonesia; emails: Gregorius.Satiabudhi@uon.edu.au, Greg@petra.ac.id; R. Chiong (corresponding author), School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia; email: Raymond.Chiong@newcastle.edu.au.

1533-5399/2023/04-ART16 $15.00
https://doi.org/10.1145/3568676

[23], create harmful effects, and eventually reduce trust in consumer reviews and undermine the effectiveness of the online market [43]. These fraudulent acts occur in response to the vital role of consumer reviews in shaping purchasing behaviour in online markets [68]; consumers trust opinions expressed in online reviews and depend on them to make decisions [4, 14, 39]. Without detection efforts, reviews may be replete with lies, fakes and deceptions, and thus completely useless—or worse [55].

The majority of studies in fake review detection follow three main approaches: those based on the content of the reviews, those based on the behaviour of the reviewers, or a combination of these [3, 29]. Content-based methods focus on the linguistic features of text such as words, **parts of speech (POS)**, n-gram, **term frequency (TF)** or other linguistic characteristics [21, 27, 55]. The behaviour-based approach focuses more on reviewers' behaviour, such as the user's identity, reviewed product, total number of reviews, ratings given, and duration of reviews [1, 3, 59]. This approach is straightforward, needs only a small number of features, and can deliver good results if properly designed. However, the behaviour-based approach is entirely dependent on additional information, such as metadata, provided by the system. In contrast, the content-based approach is independent of the system, requiring only the review text.

The content-based approach has two sub-categories. The first creates input features from the review text using **Bag of Words (BOW)**, Word2Vec, and skip-gram; thus, the input features for detection are words or terms created from the review text itself, and we call this approach textual-based featuring. The features extracted by this approach are more or less similar, mainly in the form of n-gram terms [14, 21, 27, 42, 63]. This textual-based approach is promising and is used in many studies to extract features from review texts. While independent of the system and needing only the text, this approach usually produces good results [11, 21, 27, 41, 42, 49, 55, 63, 72]. The second form of content-based method attempts to extract information and properties from the text, such as the length of the text, total words, and total sentences, in addition to linguistic characteristics of the text, such as POS, subjectivity, complexity, diversity and similarity, and sentiment analysis [12, 26, 28, 57, 65, 66, 71]. This type of content-based featuring is lightweight compared with textual-based featuring; for example, only 80 features in [12] compared with 5,000 features in [11]. This approach, too, is independent of the system since it requires only the text; however, its performance is usually relatively poor if not combined with other approaches [12].

In this study, we focus on the textual-based approach. For our experiments, we used Ott et al.'s [49] and Li et al.'s [41] datasets. These datasets are public fake review datasets used in many studies (e.g., see [12, 21, 27, 42, 55, 57, 72]). Before extracting the input features, we implemented text preprocessing such as tokenisation, stopword removal, detection of negation words, correction of elongation words, and POS lemmatisation. To extract the input features, we used the BOW method, n-gram words (from unigram to trigram words) and the count vectorised method. These methods convert a collection of text documents to a matrix of token counts. The token count features are in descending order by TF across the corpus (dataset). To detect fake reviews, we propose a novel **machine learning (ML)** ensemble model—the **Multi-type Classifier Ensemble (MtCE)**—that utilises several different types of standard (single) classifiers as its base classifier.

Ensemble classifiers are increasingly used in many different areas of study, such as text analysis [10, 39], computer security [30, 31], environmental science [62], medical research [69], electrical fault detection [47], and stock market prediction [34]. The main idea behind ensemble classifier models is to combine multiple single classifiers to produce a combination that outperforms any single classifier itself [8, 56]. In general, ensemble models can be categorised into four groups: bagging, boosting, stacked generalisation, and the mixture of experts [52]. Bagging, also called bootstrap aggregating, accumulates multiple predictors/classifiers and runs a plurality vote when predicting classes. These base classifiers are differentiated using a bootstrap technique to replicate

the learning set [5]. The **Random Forest (RF)** and Pasting Small Votes ensemble models are a variant of bagging [6, 7]. The second group of ensemble models is boosting. Similar to bagging, boosting is also an aggregation of multiple classifiers. However, instead of using bootstrap, it uses the boosting technique to train the base classifiers. This technique can convert weak learning algorithms to achieve arbitrarily high accuracy [60]. **Adaptive Boosting (AB)** [24] and **Gradient Boosting (GB)** [25] are two well-known algorithms that use this technique. In stacked generalisation [67], the classifiers are stacked to one another so that the output of some first-level base classifiers becomes the input of a second-level meta classifier. The combination of expert methods [32] combines the weighted output of several base classifiers in the consecutive combiner.

While promising, we see the potential for improvement from these ensemble classifiers; they typically use the same type of single classifier as the base classifier. To differentiate the output of each base classifier, they implement bootstrapping or boosting of the training sample inputs via bagging or boosting, stacking the classifiers, or applying different weights for their output. In contrast, our proposed MtCE combines several types of single classifiers working together as an ensemble model. As every single classifier has strengths and weaknesses in classification or detection, by combining different types of single classifiers in an ensemble, we use the strength of one classifier to 'cover' for the weakness of the other classifiers, and vice versa. The types of base classifiers ensembled in our MtCE can be customised by the user based on the problem and their experience and knowledge of this problem. To further increase the performance of our model, we implemented the bootstrap technique [5] and the method to vote on output (majority or priority) to produce the final output. For the base classifiers of our proposed ensemble model, we implemented several standard single classifiers that performed well in previous research on text mining [9–12, 16, 17], including the **Logistic Regression (LR)** [46], **Linear-kernel Support Vector Machine (LSVM)** [13, 15], **Multi-Layer Perceptron (MLP)** [58], and **Convolutional Neural Network (CNN)** [40]. In addition, two other classifiers that are usually applied as base classifiers in several ensemble models, namely the **Decision Tree (DT)** [53] and **Naïve Bayes (NB)** [44], were also used as base classifiers. The MtCE was compared with ensemble models widely used in the literature, including **Bagging Predictors (BP)** [5], RF, AB, and GB.

Theoretically, our work contributes to the artificial intelligence research domain, especially ML, by proposing a novel ensemble approach that can solve classification, detection and prediction problems. Most ensemble models in the literature either (1) have only one type of base classifier (such as RF, AB, BP and GB), or (2) have a small combination of fixed types and number of base classifiers that are designed to solve a particular problem [33, 55, 63, 64]. In contrast, our proposed model (the MtCE) provides the freedom to select the types and the number of base classifiers depending on one's requirements. Bootstrapping is included in the MtCE process to improve the model's stability and accuracy. Finally, a priority threshold approach is introduced, in addition to majority voting, to decide the final result; this priority threshold approach can improve the recall and overcome the imbalanced issue.

From a practical perspective, this work also contributes to addressing online commerce security problems—we have successfully implemented a model that can detect fake online consumer reviews with better results than previous research. Product reviews from buyers are commonly used as a guide by most consumers to make purchasing decisions on electronic commerce these days [9]. Reliable and effective detection of fake reviews will increase the trustworthiness of online commerce [10]. On the other hand, sellers use consumer reviews to evaluate the brand perception and consumers' satisfaction [22]. For this reason, maintaining the quality of product reviews is vital. Therefore, fake review detection is an urgent task and a high priority for online commerce portal providers.

Table 1. An Overview of Related Work (The Algorithms in Italics are Ensemble Models)

Author	Year	Featuring type	Algorithm
Sun et al. [63]	2016	Textual-based	*Bagging of 2 SVMs and PWCC*
Zhang et al. [71]	2016	Content- and behaviour-based	NB, SVM, DT, *RF*
Etaiwi and Naymat [21]	2017	Textual-based	NB, SVM, DT, *RF, GB*
Ren and Ji [55]	2017	Textual-based	SVM, GRNN, Recurrent Neural Network (RNN), CNN, *CNN-GRNN (Integrated)*
Dong et al. [19]	2018	Behaviour-based	*Neural Autoencoder Decision Forest*
Hazim et al.[26]	2018	Content and behaviour-based	*Generalised Boosted Regression Model (GBM) Gaussian, GBM Poisson, GBM Bernoulli, GB, AB*
Kumar et al. [39]	2018	Behaviour-based	LR, k-NN, NB, SVM, *RF, AB*
Zhang et al. [72]	2018	Textual-based	Recurrent CNN, SVM, CNN, *CNN-GRNN*
Barbado et al. [3]	2019	Behaviour-based	LR, DT, Gaussian NB (GNB), *RF, AB*
Martens & Maalej[45]	2019	Behaviour-based	DT, MLP, LSVM, RBF kernel SVM (RSVM), GNB, *RF*
Tang et al. [64]	2020	Behaviour-based	CNN, *CNN-bfGAN + SVM*
Alsubari et al. [2]	2020	Content-based	DT, *RF, AB*
Budhi et al. [11]	2021	Textual-based	LR, MLP, LSVM, *RF, AB, BP*
Budhi et al. [12]	2021	Content- and behaviour-based	DT, LR, LSVM, MLP, CNN, *RF, GB, AB, BP*
Elmogy et al. [20]	2021	Textual-based	LR, NB, k-NN, SVM, *RF*
Javed et al. [33]	2021	Textual-, content-, and behaviour-based	*Ensemble of 3 CNNs*
Shan et al. [61]	2021	Content- and behaviour-based	DT, SVM, NB, MLP, *RF*
Kumar et al. [38]	2022	Textual-, content-, and behaviour-based	Long short-term memory, Artificial Neural Network, RNN, RSVM, LR, k-NN, NB, *RF, GB, Light GB Method*

The rest of this paper is organised as follows. In the next section, we explain the design of our proposed model, the MtCE. After that, we discuss how we conducted experiments to test the performance of the MtCE, such as the datasets used, the testing framework, details of the textual-based approach, types of base classifiers, and the measurements. In Section 4, we discuss the results and analyses, and finally, we draw conclusions and outline the possibilities for future work.

2 RELATED WORK ON ENSEMBLE MODELS

The previous section provided a concise introduction to both fake review detection and ensemble models. This section discusses recent fake review detection studies from the literature (2016 to the present) that have proposed new ensemble models or implemented existing ensemble models for detection of fake reviews (see Table 1).

The objective of most previous studies has been to propose feature extraction approaches for the input of standard ML and **deep learning (DL)** methods, including their ensemble models. As discussed in Section 1, these feature extraction approaches can either be textual-based, which creates features from the words/terms of the review text itself [11, 20, 21, 72]; or content-based, which creates features from the text and extracts features from the text's information, properties, sentiment polarities, and characteristics [2]; or behaviour-based, which emphasises the behaviour of the reviewers rather than their reviews [3, 19, 39, 45].

The above approaches have also been combined to improve the detection results [12, 26, 38, 61, 71]. These studies investigated existing ML, DL, and ensemble models (i.e., AB, BP, RF, and GB) to detect fake reviews using features proposed via the combined approaches. The base classifiers of these ensemble models are of one type; for example, the RF utilises decision trees as its base classifiers, and GB utilises regression trees—both of them utilise and modify the trees in their

processes. Therefore, it is almost impossible to replace their tree-type base classifiers with others. While the type of the base classifier in some ensemble methods can be changed by design, the replacement must still be one type, e.g., it is impossible to mix more than one type of base classifier for AB and BP.

Some studies in the literature have also proposed novel ensemble models. For example, Sun et al. [63] proposed a bagging style of two SVMs and a CNN to detect fake reviews. They utilised special SVMs, namely $\text{BIGRAMS}_{\text{SVM}}$ and $\text{TRIGRAMS}_{\text{SVM}}$, to detect term features from the review text input, with a unique CNN model (**Product Word Composition Classifier (PWCC))** to capture product-related review features. The output of these three classifiers was processed in a bagging style method to produce the final detection result. Similarly, a forest of decision trees, called the Neural Autoencoder Decision Forest, was proposed by Dong et al. [19] to detect fake reviews. Their ensemble was designed by combining the Deep Neural Decision Tree (proposed by Kontschieder et al. [37]) with the Autoencoder method. An integration of several CNNs and **Gated Recurrent Neural Networks (GRNNs)** was introduced by Ren and Ji [55] in 2017. The CNNs were implemented to produce continuous sentence vectors from words, then handled by several forward and backward GRNNs to produce document representations of the reviews. A unique model of CNN, namely bfGAN, was proposed by Tang et al. [64] in 2020, and combined with the SVM for fake review detection. They showed that their model can effectively detect cold-start spam/fake reviews; the cold start problem is when a new reviewer posts a review. Javed et al. [33] proposed an ensemble of three classifiers (CNNs) for fake review detection. Each classifier detects different feature groups: a textual-based, a content-based (non-textual), and a behaviour-based group of features; a voting mechanism is then used to produce the final detection. The above-discussed approaches are similar in terms of one important aspect: they have proposed a specific ensemble model with specific types and numbers of base classifiers. In these ensemble models, the types and numbers of their base classifiers cannot be modified, since each base classifier has a specific task and purpose.

Our proposed ensemble model, the MtCE, is different. In the MtCE, diverse base classifiers can be mixed together so that the strength of one type of base classifier can overcome the weaknesses of other types of base classifiers and vice versa. The total number of base classifiers and the number of each type of base classifier can also be configured freely based on specific requirements. Our model also applies the bootstrapping technique to ensure stability and improve accuracy. For the final classification result, we introduce a priority threshold technique that prioritises a particular class in conjunction with common majority voting that is usually applied by other ensembles. This priority threshold technique can overcome the imbalanced problem when applied to the scarce/rare class as well as improve the recall in binary classification if applied to the main (positive) class.

3 THE MTCE

The MtCE is a novel ensemble of classifiers developed in light of the fact that every single classifier has its strengths and weaknesses (see Figure 1). This ensemble is proposed based on the following idea:

> *Every single classifier has been created with a different method, formulation and purpose, and thus, has strengths and weaknesses in different spots. For example, (1) for the same dataset, each classifier may correctly classify the different set of records; (2) a classifier is usually created to solve one or two problems, and not tested for the case of other problems; and (3) some classifiers perform well for smaller datasets but not for bigger ones, or vice versa. Therefore, by combining different single classifiers in one ensemble, the strengths of one classifier may 'cover' for the weaknesses of another.*

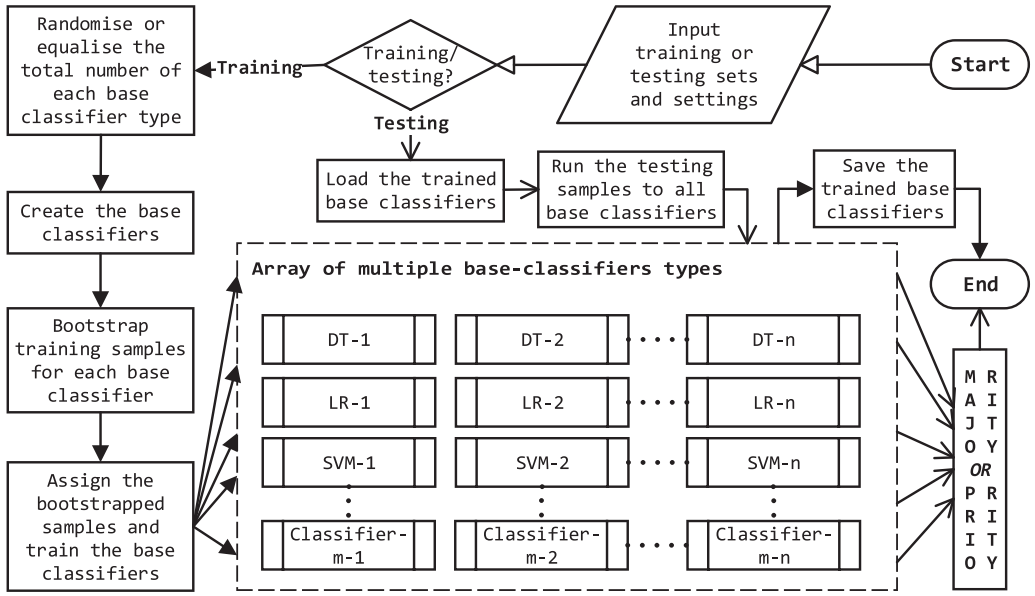

Fig. 1. Design of the MtCE.

In Figure 1, we depict two processes: the training process (the full head arrows ➔) and the testing process (the line head arrows →). All processes begin with inputting the training or testing sets and the parameter settings of the MtCE (see hollow head arrows -▷).

3.1 Base Classifiers of the MtCE

Base classifiers here include several classifier objects that are utilised within an ensemble model. These classifiers will be run jointly in a specific way (depending on the ensemble design) to produce results. These results are then processed to produce the final result of the ensemble model. For our MtCE, first the user will decide on the types of base classifiers and their total numbers. After that, for the training process, the user determines if an equal number of base classifiers will be created for each setting of each type, or if the numbers will be chosen randomly. For example, suppose the total base classifiers are 10 of three types (LR, DT, NB); an equally split setting will create 4 LR, 3 DT, and 3 NB, while for a random setting, each classifier type will be randomised from 1 to 8 (total classifiers − (total types − 1)) classifiers. While splitting the number of base classifier types equally is more sensible if the user does not know the exact nature of the problem at hand, randomising them sometimes could give better results. Moreover, with a simple modification in the implementation phase, the user can also set the exact total number of each base classifier type, if required. In this study, we did not test the effect of setting exact numbers of each type of base classifier since, for our problem, it will become similar to the randomised setting.

3.2 Bootstrap Sampling

After creating the base classifiers, each base classifier is assigned a subset of training samples based on the bootstrap sampling process. Bootstrapping the samples could improve the stability and accuracy of the model [5]. The bootstrap sampling method draws sample data repeatedly with replacement from the source (data), based on the percentage setting [5, 35]. After the training

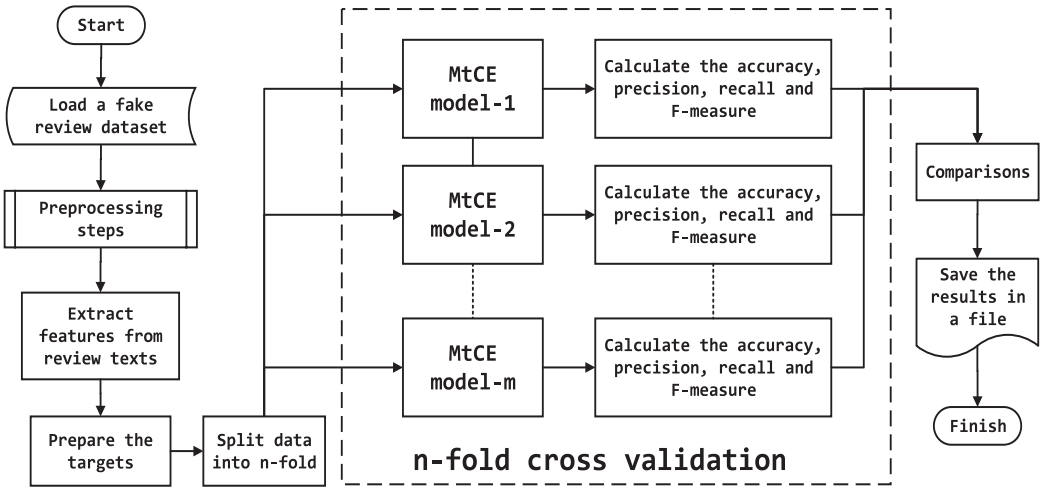

Fig. 2. Framework to test and compare the MtCE.

process for each base classifier finishes, the weights and parameters of all base classifiers are saved in a file.

3.3 Determining the Final Output

The testing process is straightforward. After inputting the testing samples and loading the previously trained base classifiers, all samples are run with the base classifiers. The final output will be determined by one of two processes: the majority vote or the priority class threshold. The majority vote chooses the majority class outputs from the results of the base classifiers. If the majority votes of all classes are equal, the final result is the smallest class in the corpus. The priority class threshold provides a final result based on the priority class chosen in the setting; that is, if the positive class is chosen as the priority, then, if the positive class outputs from base classifiers are the same as or more than the threshold, the final output is a positive class. The priority class threshold can range from absolute, which needs only one base classifier to pick the chosen class, to the maximum threshold before it becomes a majority vote (i.e., more than 50% + 1 in a binary classification problem). This priority class threshold mechanism will focus on the class that is prioritised and will increase the detection performance of this class.

4 TESTING THE PERFORMANCE OF THE MTCE

4.1 Framework for Testing

To evaluate our proposed ensemble, the MtCE, we implemented the comparison framework that we used previously to investigate classifiers for sentiment polarity detection of consumer reviews [10]. In this study, we modified the framework so that it could be applied to test the MtCE for fake review detection (see Figure 2). We used this framework to test different settings of the MtCE using similar datasets and comparing their performances. We ran the same test using several single classifiers as base classifiers of the MtCE and several well-known ensembles for comparison.

As can be seen in Figure 2, after a review text is loaded to a string, it is processed in the preprocessing subroutine to remove unnecessary words and corrections. Features are then extracted from the texts using a textual-based approach, as discussed in Section 1. After combining with the designated targets, these feature-target vectors are split into 10 folds for the **cross-validation (CV)** process. This same 10-fold setting is then trained and tested on several different

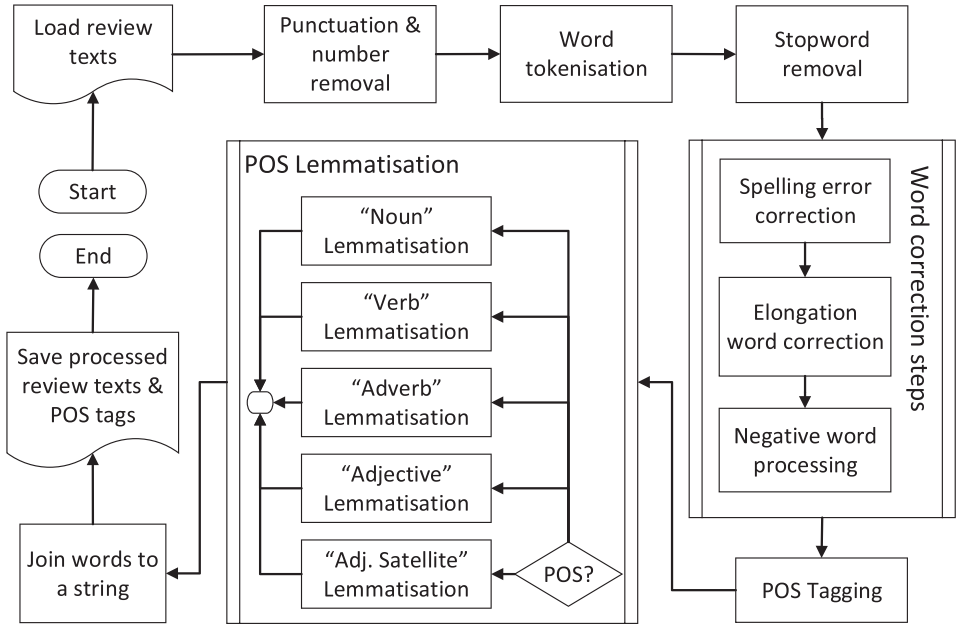

Fig. 3. Preprocessing steps [11].

combinations of MtCE models. Afterwards, the measurement results of these different MtCE models are compared to determine the best setting of the MtCE for fake review detection.

In the preprocessing steps (see Figure 3), we implemented several methods as follows:

(a) Removal of punctuation, numbers and stopwords.
(b) Correction of spelling errors, elongation words and negative words.
(c) POS tagging and lemmatisation.

Punctuation and number removal is a necessary step for a textual-based approach, since the input features of this approach are the words. Therefore, we needed to remove the non-word parts of the text reviews. Stopword is a term for words commonly used in English sentences such as 'the', 'is', 'at', 'which' and 'on'. These words can be removed from the text before it is used as a training record. These kinds of words may mislead detectors in recognising the trait words of a class, because they are often the most common words in a corpus.

Misspelled words create unnecessary issues for detection since the detector will recognise them as different words from their correct forms. Like misspelled words, elongation words such as 'yesss', 'fiiine' and 'yoouu' can also increase the diversity of words in the training example and make the classifiers harder to train. We utilised Peter Norvig's code for spelling correction [48] to correct misspelled and elongation words. Negative words have many forms, depending on the grammar, but their purpose is similar—to change a positive sentence to be negative. Therefore, we shifted all negative words to their primary form to reduce the diversity of words.

POS tagging categorises words by their syntactic function, such as noun, pronoun, adjective, verb, and adverb. This step is essential because it puts the words in their context [21] so that in the lemmatisation process, we can lemmatise the word to the correct context. Lemmatisation is a method for changing the word back to its basic form; POS lemmatisation returns the chosen word to its basic syntactic function. This step reduces the diversity of words inside the dataset and makes it easier to recognise.

Table 2. Statistics for Several Public Datasets

Dataset Author	Domain	Total Records	Fake		Genuine	
			Total	%	Total	%
Ott et al. [49]	Hotel (Various)	1600	800	50.00	800	50.00
Li et al. [41]	Doctor (Various)	558	357	63.98	201	36.02
	Hotel (Various)	1880	1080	57.45	800	42.55
	Restaurant (Various)	402	202	50.25	200	49.75

Because we extracted the input features directly from social media messages, we used the BOW feature extraction method commonly used for textual-based features [18]. This method works by decomposing the entire text into a group of singular words. Similar to previous work [11], in this study, we used it in combination with the n-gram technique to capture singular words and terms that convey one meaning but are composed of multiple words. For terms, the BOW checks for the existence of a contiguous sequence of n words from the given text sample. This study limited this checking to trigrams since sequences of more than three words are rare in real-world texts. After decomposing the text, the terms were sorted based on their frequency across the corpus.

4.2 Base Classifiers and Ensemble Classifiers for Experimental Purposes

As mentioned in Section 2, our proposed ensemble can apply multiple types of single classifiers. While in theory, any single classifier could be used as the base classifier of the MtCE, to test its performance, we investigated several combinations of ML and DL that performed well in previous research on text mining [9–12, 16, 17], which include the LR, LSVM, MLP, and CNN models. We also investigated two other classifiers, DT and NB, since these two models are commonly used as default base classifiers in some ensemble models, such as the RF [7], BP [5], and AB [24]. For comparison purposes, we tested several ensemble models, including the BP, RF, AB, and GB [25].

We built all ML classifiers and ensembles, except the CNN, using scikit-learn components [51]; the CNN was built with Keras [36] wrapped with scikit-learn components (scikit-learn does not provide a CNN component but a way to wrap DL components of Keras to be used with or within scikit-learn). To ensure the results could only be affected by our model implementation and not by modifying classifier parameters, we used the default parameters for all single classifiers used as base classifiers and the ensemble models.

4.3 Datasets

Intuitively, our proposed ensemble classifier can be applied to a wide range of problems, similar to other ML ensemble classifiers. However, to test the performance of the MtCE, we conducted experiments using four public fake review datasets (see Table 2 for additional details). The public datasets from Ott et al. [49] and Li et al. [41] were created using domain experts, such as Amazon's Mechanical Turk service (Turkers) and hotel employees, to provide false/fake reviews for some online services. Li et al.'s hotel dataset is an extension of Ott et al.'s dataset.

4.4 Cross-validation and Measurements

We evaluated the performance of the MtCE using n-fold CV, which is widely applied to evaluate the generalisation performance of an algorithm [30], to reduce bias between the dataset and the training or testing set [54] and avoid overfitting [70]. We implemented scikit-learn [51] for performance measurements of our experiments. These measurements and their respective formulas can be found in Table 3.

Table 3. Measurement Functions and Formulas

No	Name	Sklearn Function	Equation
1	Accuracy	accuracy_score()	$A(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{k=0}^{n_{samples}-1} 1(\widehat{y_k} = y_k)$ where y is the set of predicted pairs, $\hat{y}$ is the set of true pairs and $n_{samples}$ is the total number of samples.
2	Precision	precision_score()	$P(y_k, \widehat{y_k}) = \frac{tp}{tp+fp}$ where k is the set of classes, y_k is the subset of y with class k, tp is true positive, and fp is false positive.
3	Recall	recall_score()	$R(y_k, \widehat{y_k}) = \frac{tp}{tp+fn}$ where fn is false negative.
4	F-measure/F1	f1_score()	$F_1(y_k, \widehat{y_k}) = 2*\frac{P(y_k,\widehat{y_k})*R(y_k,\widehat{y_k})}{P(y_k,\widehat{y_k})+R(y_k,\widehat{y_k})}$

5 RESULTS AND DISCUSSIONS

5.1 Experimenting with the MtCE for Fake Review Detection

The first group of experiments aimed to test the performance of the MtCE for fake review detection. For its base classifiers, we implemented the CNN, LR, LSVM, and MLP, which performed well in our previous studies [9–12]. We also implemented the DT and NB, which are used as default base classifiers in many ensembles (e.g., RF, BP, and AB). The parameters of these base classifiers are the defaults of the scikit-learn components used to implement them [51]. As shown in the Appendix, we tested all possibilities from 2-, 3-, 4-, to 5-combination and presented 11 selected combinations in Table 4. Besides combining base classifiers, the other settings were fixed: total base classifiers = 15, shared equally for each type; bootstrap = 0.5; and final output = majority vote. The datasets we used in the experiments were Ott et al.'s dataset [49] and all three of Li et al.'s datasets [41]. All experiments were conducted using the 10-fold CV method. For the detailed comparison, we also tested the datasets using single classifiers as above and several well-known ensemble classifiers (RF, BP, AB, and GB). Results of the single and ensemble classifiers are provided in Table 5.

Comparing results in Table 4 with Table 5, some combinations of base classifiers in the MtCE exceed the performance of all of the base classifiers (see the bold-green scores in Table 4). For example, the accuracy of MtCE(LR-MLP) for Ott et al.'s = 88%; in comparison, in Table 5, the accuracy values of LR and MLP are 87.13% and 87.56%, respectively. However, not all combinations improved and supported each other as we hoped. A few weak classifiers degraded the performance of stronger classifiers when combined with them. This gave rise to the result that the performance of the MtCE was the average of the performance of its base classifiers; for example, the LSVM accuracy for Ott et al.'s = 85.88%, while MLP = 87.56%, and the combination of both in MtCE(LSVM-MLP) = 87.50%. This implies that the LSVM degraded the performance of MLP, and given this, rather than implementing our proposed ensemble, it would be better to use the MLP directly. However, in the same case, MtCE(LSVM-MLP) for Ott et al.'s improved the recall, from 86.21% (LSVM) and 88.37% (MLP) to 90.02%. In such cases, using our ensemble would be acceptable since the recall in binary classification/detection is the same as the accuracy of the positive class or the main class (in our problem, the fake review class). Thus, high recall means the ability of the ensemble to detect positive classes is also high.

Another finding from this set of experiments is that different datasets might need different base classifier combinations. For example, MtCE(MLP-NB) that performed best for Ott et al.'s dataset performed worst for Li et al.'s doctor dataset (increasing the recall but decreasing all other

Table 4. Results of the Combination of Base Classifiers for the MtCE Model (Selected)

Num[1]	MtCE Base Classifier Combination	Ott et al.'s Dataset[2]				Li et al.'s Dataset (Doctor)[2]			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
2	LR-MLP	**88.00**	*86.85*	**89.44**	**88.08**	84.41	*84.23*	**92.97**	88.33
5	LSVM-MLP	87.50	85.72	**90.02**	**87.74**	**86.56**	**87.05**	**93.12**	**89.88**
7	LSVM-NB	**89.19**	**88.32**	**90.43**	**89.31**	*84.75*	86.85	**90.32**	88.19
9	MLP-NB	**89.63**	**88.52**	**91.22**	**89.80**	*85.31*	*84.00*	**95.30**	*89.12*
18	LR-LSVM-NB	**89.00**	**88.68**	**89.59**	**89.07**	84.58	85.70	**91.44**	88.34
22	LSVM-MLP-NB	**89.13**	87.86	**91.00**	**89.31**	85.13	86.39	91.88	88.79
26	CNN-LR-MLP	**88.19**	**87.32**	**89.33**	**88.27**	84.40	83.75	**93.51**	88.29
35	LR-LSVM-MLP-NB	**89.19**	**88.77**	**89.95**	**89.28**	85.66	87.01	91.38	88.94
37	LR-MLP-DT-NB	**88.75**	**88.41**	89.25	**88.79**	84.06	83.85	**93.06**	88.04
47	LR-LSVM-MLP-DT-NB	**88.94**	**88.38**	**89.56**	**88.94**	84.24	84.89	92.00	88.12
50	CNN-LR-LSVM-MLP-NB	**88.88**	**88.21**	**89.83**	**88.95**	86.21	86.45	**93.18**	89.54
		Li et al.'s Dataset (Hotel)[2]				Li et al.'s Dataset (Restaurant)[2]			
2	LR-MLP	**87.34**	85.09	**94.30**	**89.45**	85.11	83.99	**88.36**	85.69
5	LSVM-MLP	86.12	84.71	**92.67**	**88.44**	**85.83**	83.98	**88.55**	**86.02**
7	LSVM-NB	**87.45**	87.09	**91.75**	**89.34**	85.35	84.51	86.33	85.21
9	MLP-NB	87.23	86.69	**91.92**	**89.17**	86.04	85.36	87.18	**86.15**
18	LR-LSVM-NB	85.96	85.20	**91.34**	88.14	**86.58**	85.71	**88.08**	**86.41**
22	LSVM-MLP-NB	86.97	86.09	**92.34**	**89.05**	85.32	83.68	**88.65**	85.75
26	CNN-LR-MLP	85.85	84.79	**91.96**	88.17	83.34	82.42	84.09	82.88
35	LR-LSVM-MLP-NB	86.70	85.12	**92.92**	88.78	82.32	81.51	84.77	82.80
37	LR-MLP-DT-NB	**87.82**	86.67	**93.26**	**89.81**	84.38	83.61	85.67	84.57
47	LR-LSVM-MLP-DT-NB	**87.55**	86.19	**93.28**	**89.58**	84.63	83.29	87.36	85.06
50	CNN-LR-LSVM-MLP-NB	86.76	85.98	**91.98**	**88.81**	83.80	82.59	86.17	83.98

[1]The number here is associated with the number in the Appendix.

[2]**Bold-green** = the MtCE result is higher than that of all base classifiers in Table 5; *Italic-red* = the MtCE result is higher than all base classifiers; Normal-black = at least one base classifier result is higher than that of the MtCE.

measurements). The best combination for Li et al.'s doctor dataset was MtCE(LSVM-MLP); for Li et al.'s hotel dataset, it was MtCE(LR-MLP-DT-NB); and for Li et al.'s restaurant dataset, it was MtCE(LR-LSVM-NB). A closer inspection of Table 5 shows that the base classifiers of MtCE's best combination for a particular dataset mostly performed well on the same dataset too, when they were used in standalone modes. For example, on Ott's MtCE(MLP-NB), the MLP and NB also performed well on Ott's dataset, although they were not as good as when combined in the MTCE. This observation is valid on other combinations for other datasets, except the DT on Li et al.'s hotel dataset, which was not performing well by itself but performed best in combination with other base classifiers in the MtCE. This indicates that, while combining good classifiers could produce better results, integrating a weak classifier sometimes could also boost the results of MtCE.

We found that the MtCE performed better than other ensembles of homogenous classifiers, as indicated by comparing the results in Table 5 (numbers 7–10) to the MtCE results in Table 4. These facts suggest that, when implemented correctly, the combination of multi-type classifiers could compensate each other's weaknesses and outperform the combination of a homogenous type of ensemble classifiers such as the RF, BP, AB, or GB. However, for the cases of RF, BP, and AB, we suspect it is also because these ensembles use the DT as their base classifiers/predictors. DT

Table 5. Results of Single and Ensemble Classifiers

Num.	Classifiers	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	CNN	79.63	81.48	77.36	79.13	67.76	70.64	88.40	77.02
2	LR	87.13	87.05	87.25	87.11	83.35	85.48	89.71	87.26
3	LSVM	85.88	85.67	86.21	85.89	83.13	86.28	87.80	86.91
4	MLP	87.56	87.05	88.37	87.66	85.67	85.89	92.94	89.16
5	DT	71.50	72.21	69.71	70.83	65.79	72.98	74.01	73.32
6	NB	88.50	87.97	89.45	88.63	87.12	90.68	88.98	89.68
7	RF	87.00	87.44	86.62	86.92	76.35	74.05	97.24	83.98
8	BP	75.00	74.23	76.79	75.41	74.19	74.50	90.21	81.45
9	AB	80.06	79.64	80.82	80.09	74.93	79.56	81.28	80.16
10	GB	82.81	83.67	81.63	82.56	76.52	77.08	90.63	82.99
		Li et al.'s Dataset (Hotel)				Li et al.'s Dataset (Restaurant)			
1	CNN	78.83	82.48	80.63	81.29	71.63	71.66	73.20	71.75
2	LR	85.59	85.50	90.25	87.77	81.59	81.03	84.52	81.74
3	LSVM	84.20	84.59	88.26	86.33	82.11	79.93	84.40	81.53
4	MLP	86.12	85.89	90.82	88.24	85.56	84.85	87.69	85.73
5	DT	68.88	73.81	71.15	72.38	60.24	60.34	63.27	61.11
6	NB	87.29	89.73	87.92	88.78	86.08	86.36	86.63	86.10
7	RF	81.97	80.78	90.28	85.17	81.34	78.31	87.86	81.88
8	BP	75.96	76.35	84.14	80.03	71.65	71.08	74.51	71.95
9	AB	79.89	80.77	84.87	82.67	70.89	70.64	70.35	70.38
10	GB	80.48	78.83	89.94	84.00	76.12	77.43	75.46	75.78

performance was not good on all datasets, which affected the performance of its ensemble variants.

While the deep learning-based CNN as a standalone classifier did not perform as well as other single classifiers, somewhat surprisingly, it performed exceptionally well when combined with other classifiers in the MtCE. This increase in accuracy was significant; for example, MtCE(CNN-LR-LSVM-MLP-NB) it is 7.93% for Li et al.'s hotel dataset and 18.45% for Li et al.'s doctor dataset. The improvement in recall was also good, with a minimum of 4.78% for Li et al.'s doctor dataset to 12.97% for Li et al.'s restaurant dataset. Therefore, the deep learning-based CNN could be combined perfectly with other single ML classifiers as base classifiers for the MtCE; it gave good results and improved all other base classifiers' performances as well. This is another strong indication that combining different classifiers with their strengths and weaknesses in our proposed MtCE can improve the overall performance. It is because the weaknesses of one base classifier can be covered by the strengths of other base classifiers (i.e., ensemble diversity).

5.2 Effects of MtCE Parameters

In this section, we discuss several input parameters of the MtCE that may affect the performance of this proposed model. All of the settings are the same as in Section 5.1 except the parameter we are testing. For these experiments, we chose five base classifier combinations of the MtCE in Table 4—that is, MtCE(LSVM-MLP), MtCE(LSVM-NB), MtCE(LR-LSVM-NB), MtCE(LR-MLP-DT-NB), and MtCE(LR-LSVM-MLP-DT-NB)—as these are the five combinations of base classifier types that have performed the best across all four datasets. We ran all experiments for parameter testing using Ott

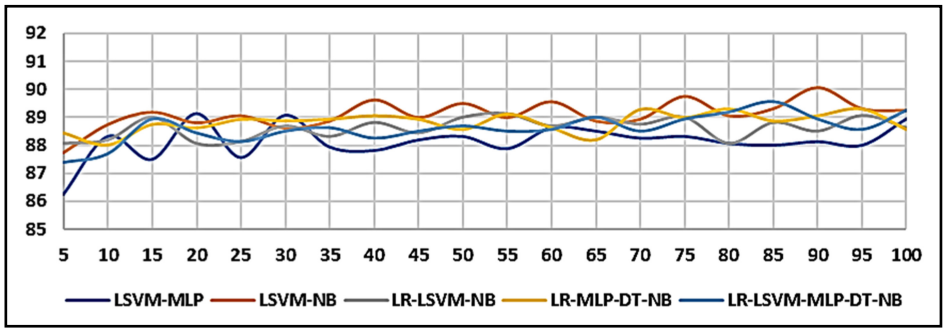

Fig. 4. The accuracy (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis).

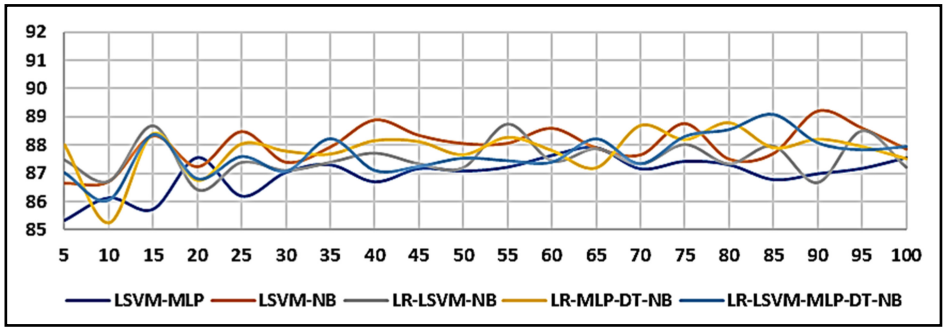

Fig. 5. The precision (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis).

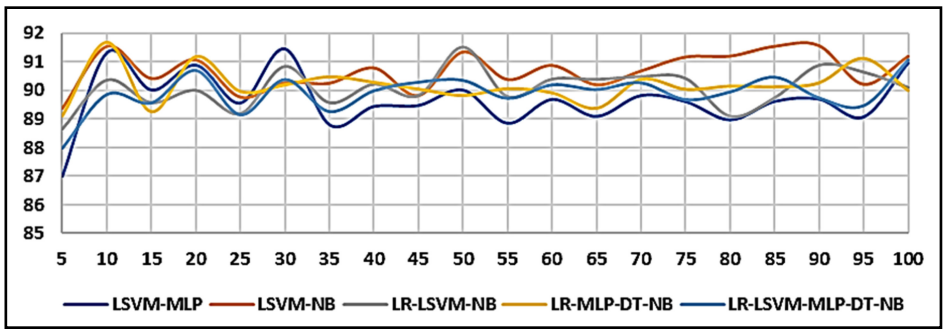

Fig. 6. The recall (y-axis, in %) of the MtCE with 5–100 total base classifiers (x-axis).

et al.'s dataset only, since these five MtCE combination settings performed well in this dataset (see the bold-green scores in Table 4). However, after finding the ideal set of parameters, we ran them with the other datasets and compared the results with our previous approach implemented on the same datasets.

5.2.1 Total Number of Base Classifiers. First, we investigated the effect of base classifiers' total numbers. Here, we ran experiments on all five combinations with the total number of base classifiers varying from 5 to 100. Results can be seen in Figure 4 for accuracy, Figure 5 for precision, Figure 6 for recall, and Figure 7 for F-measure. From these figures, we can see that accuracy, precision, recall, and F-measure increased at first, but after 20 classifiers, all scores fluctuated around a number ± 1.5%.

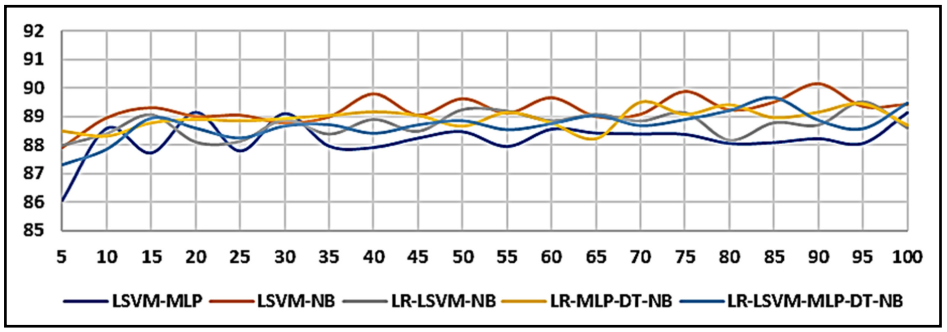

Fig. 7. The F-measure (y-axis, in %) of MtCE with 5–100 total base classifiers (x-axis).

Table 6. Results of Equal Split vs Randomly Assigned-Type of MtCE Base Classifiers on 10 Runs of the 10-fold CV Process

Base Classifier Combination	Equally Split or Randomly Assigned	Accuracy		Precision		Recall		F-measure	
		Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev	Avg.	St.Dev
LSVM-MLP	Equal	88.21	0.38	86.76	0.35	90.23	0.56	88.39	0.41
	Random	87.82	0.60	86.41	0.59	89.81	0.69	88.00	0.60
LSVM- NB	Equal	89.44	0.36	88.20	0.50	91.12	0.39	89.57	0.37
	Random	89.37	0.19	88.16	0.38	90.93	0.35	89.47	0.22
LR-LSVM-NB	Equal	88.97	0.35	88.14	0.48	90.06	0.42	89.04	0.35
	Random	88.72	0.37	87.79	0.41	89.89	0.49	88.78	0.38
LR-MLP-DT-NB	Equal	89.18	0.32	88.34	0.44	90.24	0.39	89.21	0.29
	Random	89.03	0.35	88.19	0.46	90.12	0.38	89.09	0.35
LR-LSVM-MLP-DT-NB	Equal	88.28	0.66	87.37	0.64	89.59	0.86	88.41	0.67
	Random	88.12	0.37	87.10	0.39	89.49	0.49	88.21	0.39

Accuracy fluctuated around 88.5%, precision around 88%, and recall around 90%, except for MtCE(LSVM-MLP), which was 1% lower.

Intuitively, these fluctuations after 20 classifiers mean that adding more classifiers would not improve the measurements a lot. However, since the case is only for fake review detection on Ott et al.'s dataset, we will leave it to the user decide on the number of base classifiers needed for their problem. Next, since we think a 1.5% difference is quite significant, we used the best accuracy results for each MtCE combination for the next batch of experiments. These combinations are 20 classifiers for MtCE(LSVM-MLP), 55 classifiers for MtCE(LR-LSVM-NB), 80 classifiers for MtCE(LR-MLP-DT-NB), and 85 classifiers for MtCE(LR-LSVM-MLP-DT-NB).

5.2.2 Equal Split or Randomly Assigned Base Classifier Type. To investigate the equally split vs randomly assigned base classifier type, we ran the experiments 10 times for each equally split and randomly assigned setting for each MtCE combination. The average results and their standard deviations can be seen in Table 6. This table shows that the equally split option gave slightly better average results for all measurements relative to the random option. This implies that when the total of each base classifier is equal, they support each other, and the strength of one type of classifier covers the weakness of the other type when combined in the MtCE. On the other hand, the random option could make one or two base classifiers more numerous than the others, and

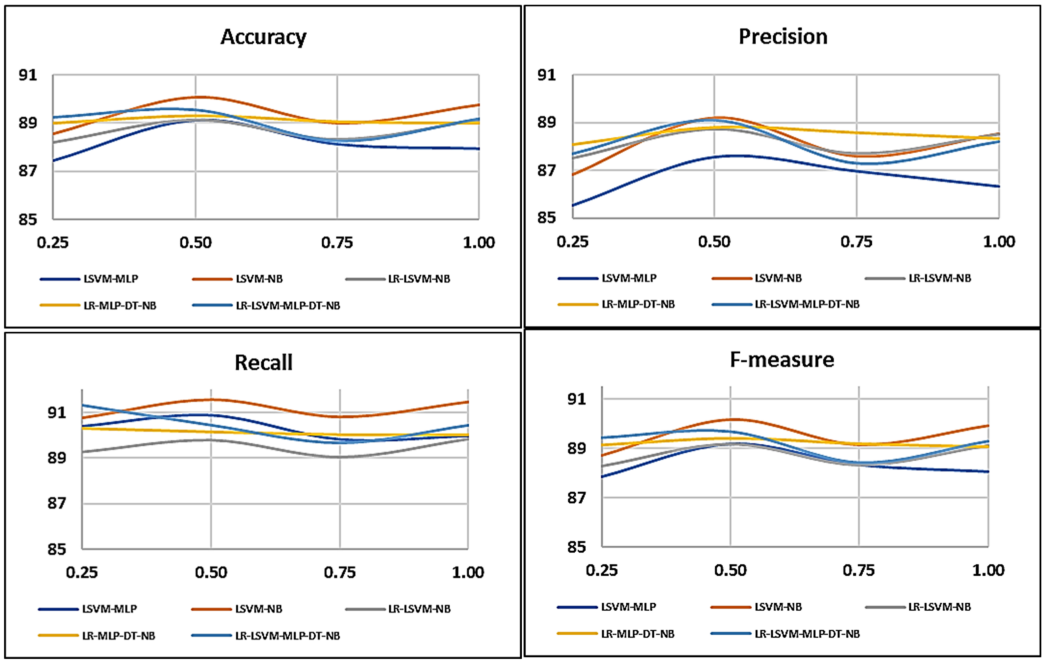

Fig. 8. The accuracy, precision, recall, and F-measure (y-axis, in %) of the MtCE with bootstrap settings from 0.25 to 1.00 (x-axis).

they dominated the detection process. The standard deviation of all measurement scores below 1 means the variation of the results from 10 runs was low and close to each other. Furthermore, we can see that the standard deviation of 2-combination, the MtCE(LSVM-MLP), on the equally split setting was lower than the random assignment. This is expected since, in an equally split setting, each base classifier total is the same for all 10 runs, while in the random option, this varies. However, the exact opposite happened in 5-combination, the MtCE(LR-LSVM-MLP-DT-NB). Here, the standard deviation of the equally split setting was higher than the random option, implying that the randomness might make the base classifiers more homogenous than when split equally between its five base classifier types. For the 3- and 4-combinations, the standard deviation scores are similar.

5.2.3 Bootstrap Effect. The subsequent experiments aimed to investigate the bootstrap parameter. We used fixed parameters as above, except the bootstrap setting ranged from 0.25 to 1 (no bootstrapping). As can be seen in Figure 8, the best accuracy, precision, and recall were often achieved when the bootstrap setting was 0.5, where every base classifier was trained using 50% of training samples randomly. This indicates that bootstrap setting 0.5 is ideal for the tested MtCE settings, i.e., the five best base-classifier combinations. Therefore, intuitively, it can be used as the default bootstrap setting of MtCE. Based on these findings, for the next set of experiments, we set the bootstrap to be 0.5.

5.2.4 Final Output: Majority Vote or Class Priority Threshold. The last parameter to test is how the output of base classifiers should be processed. There are two options in terms of transferring the output of base classifiers to the final output: majority vote and priority class. To test the effect of priority, we set the priority to the positive class (fake review class), in the range between absolute priority and 45% priority. Here, absolute priority means the final result will be recorded as

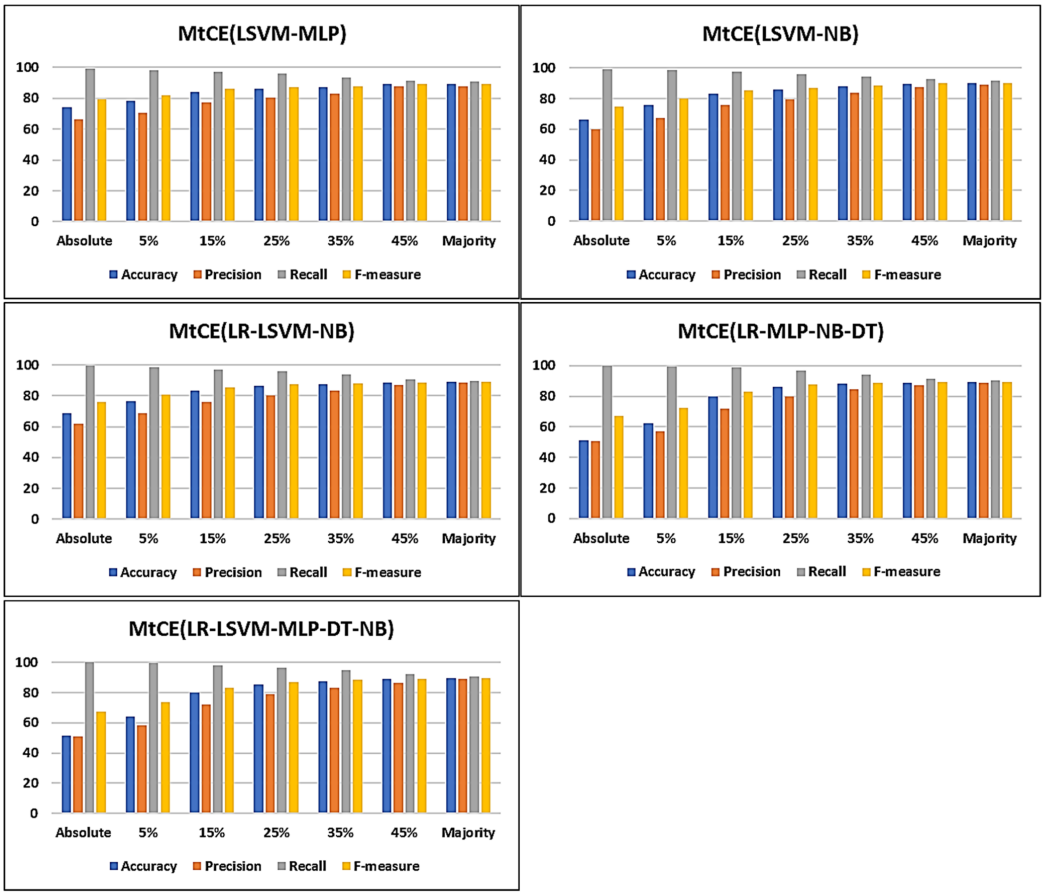

Fig. 9. Results of final target experiments, in terms of accuracy, precision, recall, and F-measure (y-axis, in %), from absolute priority to 45% priority, and majority (x-axis).

positive/fake if one or more of the base classifiers records this. Percentage priority means the final result will be recorded as positive/fake if the number of base classifiers with this result is the same or more than the percentage threshold. The results of the experiments can be seen in Figure 9.

As per Figure 9, the priority setting increases the recall but decreases other measurements. Moreover, as noted, the higher the recall, the more accurate the positive class (fake review class) detection. Thus, we can conclude that the priority setting is helpful when samples of positive class are scarce or when the focus is on anomaly detection. The key is how high we choose the percentage of priority so that the increase in recall does not greatly decrease the other measurements. For example, in Figure 9, for 25% priority of MtCE(LSVM-MLP), where recall increased by 5.05%, the accuracy, precision, and F-measure decreased by 2.75%, 7.15%, and 1.70%, respectively.

5.3 Comparison with other Studies on the Same Datasets

As commonly presented in related studies [42, 55, 57, 72], in this section, we present the MtCE's and other models' best results in light of various considerations. Many factors can influence results, such as the measurement formulas or tools, datasets used, number of records involved in experiments, and how the experiments were conducted (e.g., CV vs traditional train-test-validation

Table 7. Best Performance of the MtCE and other Studies for the Same Datasets

No.	Dataset	Description	Acc (%)	Pre (%)	Rec (%)	F1 (%)
1	Hotel (various) [49]	Ott et al. [49], SVM on positive sentiment subset	88.4	89.1	87.5	88.3
		Fusilier et al. [27], PU-L modified	-	85.2	72.8	78
		Rout et al. [57], DT	92.11	-	-	-
		Etaiwi and Naymat [21], SVM, all preproc. steps	85	51	86	-
		Zhang et al. [72], DRI-RCNN, 0.9 T.P.	-	-	-	86.59
		Budhi et al. [12], GB, over-sampling	70.62	67.58	81.29	73.8
		*MtCE(LSVM-NB, 90, Equal, 0.5, Majority)**	*90.06*	***89.19***	***91.56***	***90.16***
2	Hotel (various) [41]	Li et al. [41], OvR SAGE-Unigram	81.8	81.2	84	-
		Ren and Ji [55], Integrated, Employee/Turker subset	92.6	-	-	90.1
		Li et al. [42], SWNN	-	84.1	87	85
		Budhi et al. [12], GB, under-sampling	71.29	73.61	82.79	77.93
		*MtCE(LR-MLP-DT-NB, 15, Equal, 0.5, Majority)**	*87.82*	***86.67***	***93.26***	*89.81*
3	Restaurant (various) [41]	Li et al. [41], OvR SAGE-Unigram	81.7	84.2	81.6	-
		Ren and Ji [55], Integrated, Customer/Turker subset	86.9	-	-	86.8
		Li et al. [42], SWNN		83.3	88.2	81
		Budhi et al. [12], GB, over-sampling	72.44	71.23	75.16	73.14
		*MtCE(LR-MLP-DT-NB, 80, Equal, 0.5, Majority)**	*86.07*	***84.47***	***88.89***	*86.37*
4	Doctor (various) [41]	Li et al. [41], OvR SAGE-Unigram	74.5	77.2	70.1	-
		Ren and Ji [55], Integrated, Customer/Turker subset	76	-	-	74.1
		Li et al. [42], SWNN	-	83.7	87.6	82.9
		Budhi et al. [12], GB, over-sampling	73.35	79.44	86.94	83.02
		*MtCE(LSVM-MLP, 15, Equal, 0.5, Majority)**	***86.56***	***87.05***	***93.12***	***89.88***

*MtCE(<type of base classifiers>, <total base classifiers>, <equally split or random assign of base classifiers>, <bootstrap percentage>, <majority or priority output>).

split). Therefore, the results presented in Table 7 should be read with the following considerations in mind:

(1) We present the MtCE's and other studies' results based on the same datasets (see Table 2). We used all samples that could be processed from each dataset. Note that some studies used only a subset of the dataset or split the dataset into subsets and measured each subset differently. All of our experiments were conducted using 10-fold CV.
(2) It is impossible to ensure that all the studies used the same formulas to measure accuracy, precision, recall, and F-measure. Therefore, MtCE results were measured using widely used formulas for binary target prediction (see Table 3). All measurements were in the 'macro' average setting using scikit-learn measurement components [51].
(3) The original results from the dataset creators were used as the baseline (see the underlined description and scores in Table 7. We present only the best result for each dataset from each study.

We do not claim that the MtCE is better or worse than methods in other studies, since several factors can affect results. However, we can confidently compare the MtCE to our previous study in fake review detection [12] because the measurement formulas are similar. As is evident in Table 7, compared with our previous approach, the MtCE is superior. Accuracy improvement ranges from a minimum of 13.2% in Li et al.'s doctor dataset to a maximum of 19.4% in Ott et al.'s dataset; precision improves from 7.6% to 21.6% and recall from 6.2% to 13.7%.

6 CONCLUSION AND FUTURE WORK

From the experiments above, we can conclude that our proposed ensemble, the MtCE, was able to adequately detect fake or fraudulent reviews, which have become an acute problem on e-commerce platforms. Compared with our previous approach, the MtCE improved accuracy, precision and recall by up to 19.4%, 21.6% and 13.7%, respectively.

Not all combinations of base classifier types produced the expected result of an improvement in the performance of all base classifiers of the MtCE. In some combinations, weak classifiers dragged down the performance of stronger classifiers, especially in terms of accuracy measurements. However, when the combinations were correct, the MtCE produced better results than its base classifiers in standalone modes. It is important to note that while accuracy was not the highest, recall was the highest in many cases. This is a positive result since, in binary classification, recall is the same as accuracy of a positive case. In this case, the MtCE was able to better detect the fake review class. Comparison with well-known ensembles, such as the RF, AB, BP, and GB, showed that a correct combination of the MtCE could outperform other ensembles, proving that combining multi-type classifiers in one ensemble process performed better than combining the same classifiers, as is usually done in other ensemble methods.

We proved that DL methods such as the CNN model could be combined with ML counterparts as base classifiers to give better results than if used alone. The number of base classifiers affected performance detection; however, accuracy and other measurements oscillated after 20 base classifiers. While reducing the amount of data for the training process, the bootstrap setting could also affect the performance of the MtCE. From the experiments, we found that 50% bootstrapping gives better results than other values, including the non-bootstrap setting (bootstrap setting 100%). While the majority vote setting for the output offers a fair chance for each class to be detected, the priority threshold boosts detection of the prioritised class. This would be useful if the dataset is imbalanced or the MtCE is used to detect/classify anomalies.

Despite the extensive experimental studies presented in this paper, there remain opportunities/possibilities for further improvement. First, we are aware that multiple types of base classifiers will increase the complexity of the ensemble and increase the cost and processing time for parameter tuning. To overcome this problem, in the near future, we plan to expand on our previously proposed algorithms for ensemble parameter optimisation, namely the Multi-level **Particle Swarm Optimisation (PSO)** and its parallel version [8]. These nature-inspired algorithms are based on a low-cost PSO algorithm. They were designed to optimise the parameters of an ensemble model, such as BP and AB, choose its base classifier type, and optimise the parameters of these base classifier candidates. Other avenues for future work include investigating the implementation of the MtCE for multi-class problems, heavily imbalanced datasets, and anomaly detection.

APPENDIX

A RESULTS OF THE COMBINATION OF BASE CLASSIFIERS FOR THE MTCE MODEL

Num.	MtCE Base Classifier Combination	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	LR-LSVM	87.31	86.70	88.16	87.39	83.33	84.42	90.61	87.16
2	LR-MLP	88.00	86.85	89.44	88.08	84.41	84.23	92.97	88.33
3	LR-DT	86.69	86.96	86.48	86.67	80.82	80.96	91.62	85.91
4	LR-NB	88.56	87.47	90.07	88.72	85.29	86.51	92.21	89.00
5	LSVM-MLP	87.50	85.72	90.02	87.74	86.56	87.05	93.12	89.88
6	LSVM-DT	86.06	86.19	86.01	86.03	82.64	83.84	91.12	87.07
7	LSVM-NB	89.19	88.32	90.43	89.31	84.75	86.85	90.32	88.19
8	MLP-DT	87.81	87.80	87.28	87.50	85.13	83.71	95.45	89.01
9	MLP-NB	89.63	88.52	91.22	89.80	85.31	84.00	95.30	89.12
10	DT-NB	88.50	88.08	89.09	88.53	85.68	85.99	92.73	89.01
11	CNN-LR	85.13	85.54	84.33	84.89	82.59	83.88	90.07	86.75
12	CNN-LSVM	84.88	83.70	86.80	85.08	83.15	85.03	89.20	86.95
13	CNN-MLP	86.19	86.33	86.29	86.22	83.52	83.56	92.74	87.69
14	CNN-DT	81.31	81.96	80.53	81.15	79.59	80.82	89.31	84.75
15	CNN-NB	86.88	86.45	87.60	86.91	85.31	86.04	92.13	88.78
16	LR-LSVM-MLP	87.56	86.22	89.36	87.73	83.13	83.63	92.08	87.46
17	LR-LSVM-DT	86.50	85.61	87.68	86.56	82.61	83.56	91.25	86.89
18	LR-LSVM-NB	89.00	88.68	89.59	89.07	84.58	85.70	91.44	88.34
19	LR-MLP-DT	88.06	87.64	88.66	88.09	81.72	81.82	91.57	86.29
20	LR-MLP-NB	88.50	87.68	89.65	88.59	85.13	85.40	92.81	88.85
21	LSVM-MLP-DT	87.88	87.26	88.90	88.01	84.42	84.66	92.54	88.21
22	LSVM-MLP-NB	89.13	87.86	91.00	89.31	85.13	86.39	91.88	88.79
23	MLP-DT-NB	88.00	87.00	89.30	88.09	84.95	84.25	94.29	88.71
24	CNN-LR-DT	85.25	84.72	86.04	85.30	81.55	81.68	91.92	86.19
25	CNN-LR-LSVM	86.63	86.41	86.62	86.42	80.65	81.80	89.33	85.24
26	CNN-LR-MLP	88.19	87.32	89.33	88.27	84.40	83.75	93.51	88.29
27	CNN-LR-NB	87.50	86.60	88.59	87.52	84.58	85.68	91.33	88.23
28	CNN-LSVM-DT	84.88	85.12	84.72	84.87	83.21	84.92	90.47	87.49
29	CNN-LSVM-MLP	87.75	87.16	88.35	87.71	83.87	84.51	91.67	87.76
30	CNN-LSVM-NB	87.69	87.23	88.02	87.59	85.89	86.61	92.38	89.15
31	CNN-MLP-DT	86.50	86.55	86.25	86.29	83.70	83.51	92.99	87.82
32	CNN-MLP-NB	87.75	86.65	89.17	87.87	84.05	83.73	93.60	88.21
33	CNN-NB-DT	86.81	86.71	87.12	86.81	82.79	82.55	92.70	87.29
34	LR-LSVM-MLP-DT	87.94	87.55	88.16	87.79	82.99	82.63	92.49	87.02
35	LR-LSVM-MLP-NB	89.19	88.77	89.95	89.28	85.66	87.01	91.38	88.94
36	LR-LSVM-DT-NB	88.25	87.83	89.11	88.39	84.06	85.49	90.49	87.75
37	LR-MLP-DT-NB	88.75	88.41	89.25	88.79	84.06	83.85	93.06	88.04
38	LSVM-MLP-DT-NB	88.63	87.70	89.74	88.61	86.38	87.52	92.38	89.67
39	CNN-LR-LSVM-DT	85.94	85.82	86.07	85.87	82.44	83.18	90.71	86.65

(Continued)

Continued

Num.	MtCE Base Classifier Combination	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
40	CNN-LR-LSVM-MLP	86.88	86.69	87.33	86.90	83.17	84.29	90.92	87.29
41	CNN-LR-LSVM-NB	87.25	85.55	89.77	87.52	84.77	86.34	91.00	88.39
42	CNN-LR-MLP-DT	87.38	86.90	88.17	87.49	82.80	82.37	93.17	87.38
43	CNN-LR-MLP-NB	87.75	87.22	88.63	87.84	84.76	84.26	94.15	88.82
44	CNN-LSVM-MLP-DT	87.13	86.76	87.66	87.12	86.21	86.60	92.66	89.48
45	CNN-LSVM-MLP-NB	87.63	87.04	88.48	87.69	85.66	85.59	92.99	89.00
46	CNN-MLP-DT-NB	88.00	87.88	88.21	87.97	84.59	84.50	93.04	88.47
47	LR-LSVM-MLP-DT-NB	88.94	88.38	89.56	88.94	84.24	84.89	92.00	88.12
48	CNN-LR-LSVM-DT-NB	87.88	88.05	87.52	87.69	84.42	85.27	91.48	88.09
49	CNN-LR-LSVM-MLP-DT	87.50	87.10	88.08	87.52	82.07	82.71	91.21	86.61
50	CNN-LR-LSVM-MLP-NB	88.88	88.21	89.83	88.95	86.21	86.45	93.18	89.54
51	CNN-LR-MLP-DT-NB	87.44	86.75	87.98	87.35	85.66	85.57	93.53	89.20
52	CNN-LSVM-MLP-DT-NB	88.25	88.54	87.90	88.16	85.48	85.22	93.55	89.05
53	CNN-LR-LSVM-MLP-DT-NB	87.31	87.02	87.71	87.31	83.17	83.50	92.15	87.40
		Li et al.'s Dataset (Hotel)				Li et al.'s Dataset (Restaurant)			
1	LR-LSVM	85.37	84.23	91.73	87.75	81.13	80.49	82.52	80.59
2	LR-MLP	87.34	85.09	94.30	89.45	85.11	83.99	88.36	85.69
3	LR-DT	84.52	83.82	90.57	87.00	81.34	79.91	83.92	81.53
4	LR-NB	87.39	87.06	91.74	89.30	84.33	82.56	87.35	84.76
5	LSVM-MLP	86.12	84.71	92.67	88.44	85.83	83.98	88.55	86.02
6	LSVM-DT	83.72	83.43	89.47	86.29	81.87	81.45	83.01	81.99
7	LSVM-NB	87.45	87.09	91.75	89.34	85.35	84.51	86.33	85.21
8	MLP-DT	86.01	84.23	92.95	88.31	85.55	86.74	84.42	85.23
9	MLP-NB	87.23	86.69	91.92	89.17	86.04	85.36	87.18	86.15
10	DT-NB	85.74	86.89	88.14	87.50	85.07	85.05	84.76	84.76
11	CNN-LR	83.51	84.05	87.95	85.92	79.84	79.35	80.62	79.56
12	CNN-LSVM	82.98	83.63	87.61	85.56	81.34	83.23	80.11	81.07
13	CNN-MLP	84.68	84.71	89.66	87.03	81.82	81.39	84.25	82.20
14	CNN-DT	81.38	82.87	85.28	84.00	75.63	74.72	77.99	75.96
15	CNN-NB	84.68	86.66	86.73	86.63	83.59	82.63	86.58	84.31
16	LR-LSVM-MLP	86.54	84.99	92.96	88.73	83.33	82.29	84.50	83.27
17	LR-LSVM-DT	85.48	84.70	91.30	87.83	80.83	78.96	83.45	80.75
18	LR-LSVM-NB	85.96	85.20	91.34	88.14	86.58	85.71	88.08	86.41
19	LR-MLP-DT	87.18	85.75	93.40	89.33	84.79	84.37	85.77	84.63
20	LR-MLP-NB	87.18	86.31	92.34	89.19	84.80	83.32	86.30	84.49
21	LSVM-MLP-DT	86.33	84.69	93.17	88.69	82.35	82.10	83.87	82.35
22	LSVM-MLP-NB	86.97	86.09	92.34	89.05	85.32	83.68	88.65	85.75
23	MLP-DT-NB	86.86	86.12	92.02	88.93	84.80	84.00	85.89	84.67
24	CNN-LR-DT	84.52	83.48	91.13	87.10	79.12	77.82	80.86	79.11
25	CNN-LR-LSVM	84.57	83.82	90.67	87.04	79.35	78.24	82.94	79.95
26	CNN-LR-MLP	85.85	84.79	91.96	88.17	83.34	82.42	84.09	82.88
27	CNN-LR-NB	86.54	86.95	90.12	88.47	83.34	82.92	84.77	83.00

(Continued)

Continued

Num	MtCE Base Classifier Combination	Ott et al.'s Dataset				Li et al.'s Dataset (Doctor)			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
28	CNN-LSVM-DT	84.57	84.83	89.24	86.88	79.10	78.69	80.41	79.14
29	CNN-LSVM-MLP	85.11	84.11	91.33	87.50	83.10	82.97	83.82	83.05
30	CNN-LSVM-NB	86.33	86.77	89.95	88.26	78.50	75.48	80.49	77.64
31	CNN-MLP-DT	85.05	84.02	91.18	87.42	82.40	81.95	84.67	82.49
32	CNN-MLP-NB	87.13	86.78	91.49	89.04	84.86	87.11	85.31	86.01
33	CNN-DT-NB	85.16	85.70	88.99	87.29	82.57	84.78	80.46	82.08
34	LR-LSVM-MLP-DT	85.90	84.13	92.97	88.29	83.10	80.86	85.57	82.80
35	LR-LSVM-MLP-NB	86.70	85.12	92.92	88.78	82.32	81.51	84.77	82.80
36	LR-LSVM-DT-NB	86.28	85.79	91.29	88.41	84.35	83.62	86.91	84.73
37	LR-MLP-DT-NB	87.82	86.67	93.26	89.81	84.38	83.61	85.67	84.57
38	LSVM-MLP-DT-NB	87.18	86.13	92.67	89.20	84.57	82.77	87.86	84.89
39	CNN-LR-LSVM-DT	85.11	84.17	91.10	87.44	82.08	81.18	82.93	81.78
40	CNN-LR-LSVM-MLP	85.00	84.51	90.96	87.44	82.10	81.55	84.17	82.46
41	CNN-LR-LSVM-NB	85.85	85.19	91.30	88.09	82.88	82.45	83.32	82.62
42	CNN-LR-MLP-DT	85.43	84.25	91.93	87.87	83.54	82.08	83.21	82.24
43	CNN-LR-MLP-NB	86.54	85.87	91.66	88.62	83.62	82.77	84.95	83.54
44	CNN-LSVM-MLP-DT	85.11	84.34	91.26	87.55	80.37	80.19	82.41	80.66
45	CNN-LSVM-MLP-NB	86.76	85.74	92.15	88.82	83.35	83.36	83.37	83.02
46	CNN-MLP-DT-NB	86.91	86.48	91.43	88.87	84.32	83.74	85.90	84.59
47	LR-LSVM-MLP-NB-DT	87.55	86.19	93.28	89.58	84.63	83.29	87.36	85.06
48	CNN-LR-LSVM-DT-NB	86.28	85.77	91.39	88.44	84.03	82.07	86.59	84.07
49	CNN-LR-LSVM-MLP-DT	85.96	84.83	92.06	88.28	82.13	83.09	81.83	81.64
50	CNN-LR-LSVM-MLP-NB	86.76	85.98	91.98	88.81	83.80	82.59	86.17	83.98
51	CNN-LR-MLP-DT-NB	86.06	85.09	91.85	88.31	84.30	84.98	83.43	84.04
52	CNN-LSVM-MLP-DT-NB	85.74	85.20	91.13	88.01	83.59	82.36	85.31	83.56
53	CNN-LR-LSVM-MLP-DT-NB	85.32	85.38	89.80	87.50	83.09	82.68	83.40	82.87

REFERENCES

[1] A. U. Akram, H. U. Khan, S. Iqbal, T. Iqbal, E. U. Munir, and M. Shafi. 2018. Finding rotten eggs: A review spam detection model using diverse feature sets. *KSII Transactions on Internet and Information Systems* 12, 10 (2018). DOI: http://dx.doi.org/10.3837/tiis.2018.10.026

[2] S. N. Alsubari, M. B. Shelke, and S. N. Deshmukh. 2020. Fake reviews identification based on deep computational linguistic features. *International Journal of Advanced Science and Technology* 29, 8s (2020), 3846–3856.

[3] R. Barbado, O. Araque, and C. A. Iglesias. 2019. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management* 56, 4 (2019), 1234–1244. DOI: http://dx.doi.org/10.1016/j.ipm.2019.03.002

[4] L. Bing, T.-L. Wong, and W. Lam. 2016. Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. *ACM Transactions on Internet Technology* 16, 2 (2016), 1–17. DOI: http://dx.doi.org/10.1145/2857054

[5] L. Breiman. 1996. Bagging predictors. *Machine Learning* 24, 2 (1996), 123–140. DOI: https://doi.org/10.1007/bf00058655

[6] L. Breiman. 1999. Pasting small votes for classification in large databases and on-line. *Machine Learning* 36, 1 (1999), 85–103. DOI: http://dx.doi.org/10.1023/A:1007563306331

[7] L. Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32. DOI: https://doi.org/10.1023/a:1010933404324

[8] G. S. Budhi, R. Chiong, and S. Dhakal. 2020. Multi-level particle swarm optimisation and its parallel version for parameter optimisation of ensemble models: A case of sentiment polarity prediction. *Cluster Computing* 23 (2020), 3371–3386. DOI: https://doi.org/10.1007/s10586-020-03093-3

[9] G. S. Budhi, R. Chiong, I. Pranata, and Z. Hu. 2017. Predicting rating polarity through automatic classification of review texts. In *Proceedings of the 2017 IEEE Conference on Big Data and Analytics (ICBDA)*. Kuching, Malaysia, 19–24. DOI: https://doi.org/10.1109/ICBDAA.2017.8284101

[10] G. S. Budhi, R. Chiong, I. Pranata, and Z. Hu. 2021. Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis. *Archives of Computational Methods in Engineering* 28 (2021), 2543–2566. DOI: https://doi.org/10.1007/s11831-020-09464-8

[11] G. S. Budhi, R. Chiong, and Z. Wang. 2021. Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications* 80 (2021), 13079–13097. DOI: https://doi.org/10.1007/s11042-020-10299-5

[12] G. S. Budhi, R. Chiong, Z. Wang, and S. Dhakal. 2021. Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews. *Electronic Commerce Research and Applications* 47 (2021), 101048. DOI: https://doi.org/10.1016/j.elerap.2021.101048

[13] C. Campbell and Y. Ying. 2011. *Learning with Support Vector Machines*. Morgan & Claypool.

[14] E. F. Cardoso, R. M. Silva, and T. A. Almeida. 2018. Towards automatic filtering of fake reviews. *Neurocomputing* 309 (2018), 106–116. DOI: http://dx.doi.org/10.1016/j.neucom.2018.04.074

[15] C. C. Chang and C. J. Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3 (2011), 1–27. DOI: https://doi.org/10.1145/1961189.1961199

[16] R. Chiong, G. S. Budhi, and S. Dhakal. 2021. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems* 36, 6 (2021), 99–105. DOI: https://doi.org/10.1109/MIS.2021.3093660

[17] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong. 2021. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine* 135 (2021), 104499. DOI: https://doi.org/10.1016/j.compbiomed.2021.104499

[18] X. Deng, Y. Li, J. Weng, and J. Zhang. 2019. Feature selection for text classification: A review. *Multimedia Tools and Applications* 78, 3 (2019), 3797–3816. DOI: https://doi.org/10.1007/s11042-018-6083-5

[19] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning. 2018. Opinion fraud detection via neural autoencoder decision forest. *Pattern Recognition Letters* 132 (2018), 21–29. DOI: http://dx.doi.org/10.1016/j.patrec.2018.07.013

[20] A. M. Elmogy, U. Tariq, and A. I. A. Mohammed. 2021. Fake reviews detection using supervised machine learning. *International Journal of Advanced Computer Science and Applications* 12, 1 (2021), 601. DOI: https://dx.doi.org/10.14569/IJACSA.2021.0120169

[21] W. Etaiwi and G. Naymat. 2017. The impact of applying different preprocessing steps on review spam detection. *Procedia Computer Science* 113 (2017), 273–279.

[22] A. Felbermayr and A. Nanopoulos. 2016. The role of emotions for the perceived usefulness in online customer reviews. *Journal of Interactive Marketing* 36 (2016), 60–76. DOI: http://dx.doi.org/10.1016/j.intmar.2016.05.004

[23] V. W. Feng and G. Hirst. 2013. Detecting deceptive opinions with profile compatibility. In *Proceedings of International Joint Conference on Natural Language Processing*. Nagoya, Japan, 338–346.

[24] Y. Freund and R. E. Schapire. 1997. A decision-theoretic generalisation of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 1 (1997), 119–139. DOI: https://doi.org/10.1006/jcss.1997.1504

[25] J. H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232

[26] M. Hazim, N. B. Anuar, M. F. A. B. Razak, and N. A. Abdullah. 2018. Detecting opinion spams through supervised boosting approach. *PLoS ONE* 13, 6 (2018), e0198884. DOI: http://dx.doi.org/10.1371/journal.pone.0198884

[27] D. Hernández Fusilier, M. Montes-Y-Gómez, P. Rosso, and R. Guzmán Cabrera. 2015. Detecting positive and negative deceptive opinions using PU-learning. *Information Processing & Management* 51, 4 (2015), 433–443. DOI: http://dx.doi.org/10.1016/j.ipm.2014.11.001

[28] A. Heydari, M. Tavakoli, and N. Salim. 2016. Detection of fake opinions using time series. *Expert Systems with Applications* 58 (2016), 83–92. DOI: http://dx.doi.org/10.1016/j.eswa.2016.03.020

[29] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari. 2015. Detection of review spam: A survey. *Expert Systems with Applications* 42, 7 (2015), 3634–3642. DOI: http://dx.doi.org/10.1016/j.eswa.2014.12.029

[30] Z. Hu, R. Chiong, I. Pranata, Y. Bao, and Y. Lin. 2019. Malicious web domain identification using online credibility and performance data by considering the class imbalance issue. *Industrial Management & Data Systems* 119, 3 (2019), 676–696. DOI: https://doi.org/10.1108/IMDS-02-2018-0072

[31] R. Chiong, Z. Wang, Z. Fan, and S. Dhakal. 2022. A fuzzy-based ensemble model for improving malicious web domain identification. *Expert Systems with Applications* 204 (2022), 117243. DOI: https://doi.org/10.1016/j.eswa.2022.117243

[32] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation* 3, 1 (1991), 79–87. DOI: https://doi.org/10.1162/neco.1991.3.1.79

[33] M. S. Javed, H. Majeed, H. Mujtaba, and M. O. Beg. 2021. Fake reviews classification using deep learning ensemble of shallow convolutions. *Journal of Computational Social Science* 4, 2 (2021), 883–902. DOI: http://dx.doi.org/10.1007/s42001-021-00114-y

[34] R. Chiong, Z. Fan, Z. Hu, and S. Dhakal. 2022. A novel ensemble learning approach for stock market prediction based on sentiment analysis and the sliding window method. *IEEE Transactions on Computational Social Systems*. DOI: http://dx.doi.org/10.1109/TCSS.2022.3182375

[35] R. W. Johnson. 2001. An introduction to the bootstrap. *Teaching Statistics* 23, 2 (2001), 49–54. DOI: https://doi.org/10.1111/1467-9639.00050

[36] Keras. 2019. Keras: The Python Deep Learning library.

[37] P. Kontschieder, M. Fiterau, A. Criminisi, and S. R. Bul'O. 2015. Deep neural decision forests. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 1467–1475.

[38] A. Kumar, R. D. Gopal, R. Shankar, and K. H. Tan. 2022. Fraudulent review detection model focusing on emotional expressions and explicit aspects: Investigating the potential of feature engineering. *Decision Support Systems* 155 (2022), 113728. DOI: http://dx.doi.org/10.1016/j.dss.2021.113728

[39] N. Kumar, D. Venugopal, L. Qiu, and S. Kumar. 2018. Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems* 35, 1 (2018), 350–380. DOI: http://dx.doi.org/10.1080/07421222.2018.1440758

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 11 (1998), 2278–2324. DOI: https://doi.org/10.1109/5.726791

[41] J. Li, M. Ott, C. Cardie, and E. Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD, USA, 1566–1576.

[42] L. Li, B. Qin, W. Ren, and T. Liu. 2017. Document representation and feature combination for deceptive spam review detection. *Neurocomputing* 254 (2017), 33–41. DOI: http://dx.doi.org/10.1016/j.neucom.2016.10.080

[43] J. Malbon. 2013. Taking fake online consumer reviews seriously. *Journal of Consumer Policy* 36, 2 (2013), 139–157. DOI: http://dx.doi.org/10.1007/s10603-012-9216-7

[44] C. D. Manning, P. Raghavan, and H. Schuetze. 2008. Naïve Bayes text classification. In *Introduction to Information Retrieval* Cambridge University Press, 234–265.

[45] D. Martens and W. Maalej. 2019. Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering* 24, 6 (2019), 3316–3355. DOI: http://dx.doi.org/10.1007/s10664-019-09706-9

[46] S. Menard. 2010. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE, Los Angeles, CA.

[47] Y. Mohammadi, A. Salarpour, and R. Chouhy Leborgne. 2021. Comprehensive strategy for classification of voltage sags source location using optimal feature selection applied to support vector machine and ensemble techniques. *International Journal of Electrical Power & Energy Systems* 124 (2021), 106363. DOI: http://dx.doi.org/10.1016/j.ijepes.2020.106363

[48] P. Norvig. 2016. How to write a spelling corrector.

[49] M. Ott, C. Cardie, and J. T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia, USA, 497–501.

[50] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, Oregon, 309–319.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830.

[52] R. Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 3 (2006), 21–45. DOI: http://dx.doi.org/10.1109/MCAS.2006.1688199

[53] J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106. DOI: https://doi.org/10.1007/bf00116251

[54] Q. Ren, M. Li, and S. Han. 2019. Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: A comparative study from multiple perspectives. *Big Earth Data* 3, 1 (2019), 8–25. DOI: https://doi.org/10.1080/20964471.2019.1572452

[55] Y. Ren and D. Ji. 2017. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences* 385–386, 213–224. DOI: http://dx.doi.org/10.1016/j.ins.2017.01.015

[56] Y. Ren, L. Zhang, and P. N. Suganthan. 2016. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine* 11, 1 (2016), 41–53. DOI: http://dx.doi.org/10.1109/mci.2015.2471235

[57] J. K. Rout, S. Singh, S. K. Jena, and S. Bakshi. 2016. Deceptive review detection using labeled and unlabeled data. *Multimedia Tools and Applications* 76, 3 (2016), 3187–3211. DOI: http://dx.doi.org/10.1007/s11042-016-3819-y

[58] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, 318–362.

[59] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang. 2015. Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications* 42, 22 (2015), 8650–8657. DOI: http://dx.doi.org/10.1016/j.eswa.2015.07.019

[60] R. E. Schapire. 1990. The strength of weak learnability. *Machine Learning* 5, 2 (1990), 197–227. DOI: http://dx.doi.org/10.1007/BF00116037

[61] G. Shan, L. Zhou, and D. Zhang. 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems* 144 (2021), 113513. DOI: http://dx.doi.org/10.1016/j.dss.2021.113513

[62] J. Shin, S. Yoon, Y. Kim, T. Kim, B. Go, and Y. Cha. 2021. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics* 61 (2021), 101202. DOI: http://dx.doi.org/10.1016/j.ecoinf.2020.101202

[63] C. Sun, Q. Du, and G. Tian. 2016. Exploiting product related review features for fake review detection. *Mathematical Problems in Engineering 2016*. 1–7. DOI: http://dx.doi.org/10.1155/2016/4935792

[64] X. Tang, T. Qian, and Z. You. 2020. Generating behavior features for cold-start spam review detection with adversarial learning. *Information Sciences* 526 (2020), 274–288. DOI: http://dx.doi.org/10.1016/j.ins.2020.03.063

[65] E. D. Wahyuni and A. Djunaidy. 2016. Fake review detection from a product review using modified method of iterative computation framework. In *Proceedings of MATEC Web of Conferences* 58, 03003. DOI: http://dx.doi.org/10.1051/matec

[66] X. Wang, K. L. S. He, and J. Zhao. 2016. Learning to represent review with tensor decomposition for spam detection. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Texas, USA, 866–875.

[67] D. H. Wolpert. 1992. Stacked generalisation. *Neural Networks* 5, 2 (1992), 241–259. DOI: https://doi.org/10.1016/S0893-6080(05)80023-1

[68] Y. Wu, E. W. T. Ngai, P. Wu, and C. Wu. 2020. Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems* 132 (2020), 113280. DOI: http://dx.doi.org/10.1016/j.dss.2020.113280

[69] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik. 2017. Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Transactions on Medical Imaging* 36, 3 (2017), 849–858. DOI: http://dx.doi.org/10.1109/TMI.2016.2633551

[70] T. Xiong, Y. Bao, Z. Hu, and R. Chiong. 2015. Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. *Information Sciences* 305 (2015), 77–92. DOI: https://doi.org/10.1016/j.ins.2015.01.029

[71] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic. 2016. What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. *Journal of Management Information Systems* 33, 2 (2016), 456–481. DOI: https://doi.org/10.1080/07421222.2016.1205907

[72] W. Zhang, Y. Du, T. Yoshida, and Q. Wang. 2018. DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management* 54, 4 (2018), 576–592. DOI: http://dx.doi.org/10.1016/j.ipm.2018.03.007

Received 6 November 2021; revised 12 July 2022; accepted 29 August 2022